Estimation of Articulatory Parameters from the Acoustic Speech Signal

Alexander Sepulveda

Universidad Nacional de Colombia–sede Manizales
Facultad de Ingeniería y Arquitectura
Departamento de Ingenierías Eléctrica, Electrónica y Computación
Manizales, Colombia
2012

Estimación de Parámetros Articulatorios a partir de la Señal de Voz

Alexander Sepulveda

Tesis presentada como requisito parcial para optar al título de:
Doctor en Ingeniería - Automática

Director:
Germán Castellanos-Domínguez

Línea de Investigación:
Automática-Procesamiento Digital de Señales
Grupo de Investigación:
Signal Processing and Recognition Group

Universidad Nacional de Colombia-sede Manizales
Facultad de Ingeniería y Arquitectura
Departamento de Ingenierías Eléctrica, Electrónica y Computación
Manizales, Colombia
2012

To my son.

I lovingly dedicate this thesis to my son, who gave me the time and the motivation to finish it.

A man who carries a cat by the tail learns something he can learn in no other way.

–Mark Twain–

El presente trabajo fue financiado por el Departamento Administrativo de Ciencia, Tecnología e Innovación de Colombia-**COLCIENCIAS**, mediante la convocatoria *Apoyo a la comunidad científica nacional, a través de los programas de doctorado nacionales-2006.*

Agradecimientos

Esta tesis doctoral, si bien ha requerido de esfuerzo y dedicación, no hubiese sido posible realizarla sin la cooperación desinteresada de otras personas. En sentido estricto, el número de personas a las cuales debo agradecimientos en enorme ya que, como dijo Erich Fromm, una vez dos seres humanos han interactuado, ninguno de los dos es el mismo; sin embargo, debido a la falta de espacio, solo puedo mencionar unas pocas personas, los demás ... que me perdonen.

En primera medida quiero agradecer al prof. Germán Castellanos, el supervisor del presente trabajo de investigacion, por confiar en mi, por su paciencia y sugerencias constructivas durante todo este tiempo. Además quiero agradecer a los profesores Yves Laprie (INRIA-LORIA, Francia) y Rodrigo Capobianco (Universidad de São Paulo, Brasil), quienes hicieron las veces de tutor durante mis estancias en Francia y Brasil, respectivamente. Ellos me prestaron su ayuda en momentos criciales de la presente tesis.

Agradezco a mis compañeros de grupo de investigacion en la U. Nacional, con quienes uno comparte conocimiento, puntos de vista y hasta programas. Agradecimientos para Genaro, Jorge, Juan Bernando, Carlos Guerrero, Julian y Luis. Agradezco además a aquellos otros compañeros con las cuales compartí momentos especiales en Manizales: Andrés, Johana, Victoria Eugenia, Patricia, Alexander, Kerly, Ivan, Diana, Sebastian, Bertha, Valentina, Germán B. y Jose Luis; y a todas aquellas personas que de una otra forma evitaron que el doctarado fuera más solitario.

Especiales agradecimientos para Farid Feiz y Asterios Toutios, quienes me colaboraron bastante durante mi estancia en Francia. Quedaré siempre agradecido con Catalina, para quien siempre he guardado un afecto especial, aunque ella no lo sepa. Agradezco además a Diego Patiño, Franz Stoten y a Julie Busset, por los momentos compartidos en Francia.

Mis agradecimientos para los profesores del DIEEC, de quienes he recibido siempre su apoyo; y por supuesto a Sonia y Maria Clemencia por su paciencia para con mis viscitudes en lo que a documentos y trámites se refiere.

Agradezco a mi familia por ese apoyo silencioso y fuerte; en especial a mi hijo, a quien dedico el presente trabajo. A mis padres y hermanos, al igual que yo, ahora estamos más tranquilos. Y por supuesto a O. Lucía por su compañía incondicional.

Agradezco además a aquellas personas que en algún momento de mi vida me ayudaron de forma definitiva, a la profesora Elsa en mi primaria, los profesores Henry C., Manuel S., Marina O. y Carmen Y. en mi secundaria. Especiales agradecimiemtos a Israel Gonzáles, mi compañero de ajedrez en la adolecencia; además, me prestó un libro del cual aun guardo los apuntes hechos a mano. Estos apuntes los leí de vez en cuando durante la realización de la tesis, y en particular en aquellos momentos más difíciles.

Declaración

Me permito afirmar que he realizado la presente tesis de manera autónoma y con la única ayuda de los medios permitidos y no diferentes a los mencionados en la propia tesis. Todos los pasajes que se han tomado de manera textual o figurativa de textos publicados y no publicados, los he reconocido en el presente trabajo. Ninguna parte del presente trabajo se ha empleado en ningún otro tipo de tesis.

Bogotá, D.C., Junio de 2012

_____

(Alexander Sepulveda)

Abstract

The articulatory inversion, if it could be done in a practical way, would have several applications; namely: in speech therapy applications and language learning systems for training pronunciation; to reduce problems caused by coarticulation and noise in automatic speech recognition systems; among other applications. Due to the range of applications of articulatory inversion, it has captivated the attention of speech scientist during several decades. However, the available human articulatory data were scarce. On the other hand, technologies such as electromagnetic articulography have made the measurement of human articulation during speech be more accessible. In order to take advantage of human articulation measurements, several methods have been tested; e.g., artificial neural networks, hidden Markov models, Gaussian mixture models, among others. But, less attention has been put into the influence of the kind of acoustic features used in those methods.

The aim of this thesis is to show the importance of selecting the acoustic input features in those tasks related to the inference of articulators movements during the speech signal production. Analyzed parameters include: the formants, time-frequency representation using the wavelet transform as well as time-frequency representation using filter banks in *mel* scale. In the case of the time-frequency representations, those characteristics localized in time and frequency that allow a more accurate estimate of the vocal tract shape are considered.

It is found that there exist some actions that improve the performance of acoustic-to-articulatory mapping systems, namely: 1) using those time-frequency features best related to articulators movement from the perspective of non-linear statistical correlation, which we call maps of relevant time-frequency features; and, 2) including features intrinsically related to the vocal-tract resonance frequencies in the input set of features representing the speech signal.

Additionally, in case of fricative sounds, it is shown in present study that the maps of relevant time-frequency features are also useful for speaker-independient tasks; then, the same proposed approach could be used for the further development of a multi-speaker acoustic-to-articulatory mapping. Once obtained the multispeaker articulatory inversion system, it could be used in speech therapy related tasks, particularly in speech training for the cleft palate children. Another potential application are computer-based language learning systems.

Keywords: articulatory inversion, speech production modelling, wavelet transform, articulatory parameters, articulatory synthesizer, Gaussian mixture models, artificial neural networks.

Resumen

La inversión articulatoria, si existiese una manera práctica de realizarla, tendría varias aplicaciones, por ejemplo: en aplicaciones de terapia del habla y sistemas de aprendizaje de idiomas para el entrenamiento de la pronunciación, para reducir los problemas causados por la coarticulación y el ruido en sistemas automáticos de reconocimiento de voz, entre otras aplicaciones. Debido al rango de aplicaciones de la inversión articulatoria, esta ha cautivado la atención de científicos del habla durante varias décadas. Sin embargo, los datos articulatorios reales disponibles eran escasos. Por otra parte, las tecnologías como la articulografía electromagnética han hecho que la medición de la articulación humana durante el habla sea más accesible. Con el fin de aprovechar la disponibilidad mediciones del mecanismo articulatorio varios métodos han sido probados. Por ejemplo, redes neuronales artificiales, modelos ocultos de Markov, modelos de mezclas gaussianas, entre otros. Pero, poca atención se le ha prestado a la influencia del tipo de características acústicas utilizadas en estos métodos.

La presente tesis tiene por objetivo principal el mostrar la importancia que tiene la selección de los parámetros acústicos, los cuales son usados para representar la voz, en tareas de inversión articulatoria; es decir, en tareas relacionadas con la inferencia de la posición de los articuladores durante la producción de la misma señal de voz. Dentro de los parámatros acústicos analizados se mencionan: los formantes, representación de tiempo-frecuencia por medio de la transformada *wavelet* y mediante banco de filtros en la escala *mel*. Para el caso de las representaciones de tiempo-frecuencia se buscan aquellas características localizadas en tiempo y frecuencia que permiten una estimación más precisa de la forma del tracto vocal.

A modo de resultado se encuentra que existen dos acciones que mejoran la estimación de la posición de los articuladores, a saber: 1) usar caractarísticas de tiempo-frecuencia que desde el punto de vista de la correlación estadística no-lineal están mejor relacionadas con las trayectorias de los movimientos articulatorios; y, 2) incluir dentro del conjunto de

representación de la señal de voz parámetros intrínsicamente relacionados con las frecuencias de resonancia del tracto vocal.

Hasta donde se conoce, aún no se ha desarrollado un sistema para la inversión articulatoria independiente del hablante. Sin embargo, en el presente trabajo se muestra que los mismos mapas de características relevantes de tiempo-frecuencia pueden ser utilizadas para la realización de la inversión articulatoria independiente del hablante sobre consonantes fricativas.

A modo de trabajo futuro se plantea desarrollar un sistema de inversión articulatoria independiente del hablante basado en mapas de relevancia, los cuales serían obtenidos para varias categorías fonéticas. Se tiene planeado, una vez hecho esto, utilizar los resultados para el desarrollo de sistemas de terapia de la voz y en el aprendizaje de idiomas.

Palabras clave: Inversión articulatoria, modelado del mecanismo de producción del habla, transformada ondita, parámetros acústicos, sintetizador articulatorio, modelos de mezclas gaussianas, redes neuronales.

Content

Abstract xi

List of symbols xvi

1. Introduction 1

1.1. Motivation . . . 1
1.2. Review of articulatory inversion . . . 3
1.2.1. Measurement of the articulatory phenomenon . . . 4
1.2.2. Inversion methods . . . 7
1.2.3. Inputs to inversion: representation of the speech signal . . . 13
1.3. Problem statement . . . 16
1.4. Thesis outline . . . 18
1.5. Publications . . . 20

2. Parametrization of the speech signal 22

2.1. The nature of speech . . . 22
2.1.1. Production of speech signal . . . 22
2.1.2. Introduction to the acoustics of phonemes . . . 24
2.1.3. Critical articulators. . . . 26
2.2. Speech signal representation . . . 27
2.2.1. Cepstrum . . . 27
2.2.2. Formants . . . 29
2.2.3. Representation based on wavelet packet transform . . . 31
2.2.4. Filter-banks . . . 36

3. Proposed methods for articulatory inversion 39

3.1. TF relevant features for inversion . . . 39
3.1.1. Measures of statistical association . . . 39
3.1.2. Regression by using Gaussian mixture models . . . 42

3.2. Vocal tract modeling 45
3.2.1. Acoustic modeling 46
3.2.2. Articulatory model 50
3.3. Weighted cepstral distance learning for accessing articulatory codebooks . . . 52
3.3.1. Cepstral distance measure 52
3.3.2. Cost function 54
3.3.3. Optimization of the cost function 56

4. Experimental setup **59**
4.1. Testing cepstral distance measures 59
4.2. Testing the contribution of formants on acoustic-to-articulatory mapping systems 61
4.3. Testing the relevant maps of TF features 63
4.3.1. Articulatory data 63
4.3.2. TF relevant features for inversion 65
4.3.3. Measures of performance 67

5. Role of formants on articulatory inversion **67**
5.1. Resulting cepstral distance 67
5.2. Contribution of formants on inversion systems based on neural network . . . 71
5.3. Discussion 72

6. Maps of relevant TF features as a mean for improving performance of the acoustic-to-articul
6.1. On estimating time-frequency maps of relevant features 80
6.2. Assesment of Kendall coefficient for measuring the statistical dependence between articulators
6.3. Relevant acoustic-to-articulatory maps using Gaussian mixture regression . . 89
6.4. TF relevant features for subject-independent acoustic-to-articulatory mapping of fricatives 93
6.4.1. Speech signal representation 96
6.4.2. Inversion of fricatives using relevant TF features 97
6.4.3. Subject-independent inversion of fricatives using relevant TF features and VTLN100
6.5. Discussions 102

7. Conclusions **107**

References **110**

List of symbols

Symbols

Symbol	
$\boldsymbol{a}$	approximation wavelet coefficients
$\boldsymbol{c}$	cepstrum
d	time lag
$\mathrm{d}(\cdot,\cdot)$	distance measure
$\boldsymbol{d}$	detail wavelet coefficients
$\boldsymbol{E}$	excitation spectra
F_i	i_{th} formant
f_k	index of the filter-bank
$\boldsymbol{F}$	vector of formant values
f	frequency
E_{RMS}	root mean square error
$\boldsymbol{H}$	vocal-tract spectra
h_0	approximation wavelet filter
h_1	detail wavelet filter
$I(\cdot,\cdot)$	mutual information measure
li	lower incisors
ll	lower lip
Λ	relation between Maeda's parameters and vocal tract form

Symbol	
$\boldsymbol{\mu}$	mean vector
N_t	number of frames forming the context-window analysis
N_f	number of filters in the filter-bank
N_e	number of samples in the signal $\boldsymbol{s}(t)$
N_r	number of real speech frames
N_s	quantity of codebook entries
N_c	quantity of articulatory channels
$P(\cdot)$	Probability
$\psi_{\cdot,\cdot}(t)$	mother wavelet function
ρ	linear correlation vale
$\boldsymbol{S}(\cdot)$	discrete spectrum
$\boldsymbol{\Sigma}$	covariance matrix
tt	tongue tip
tb	tongue body
td	tongue dorsum
t	time position
t_a, t_b	time lags envolving context-window analysis
τ	Kendall value
$T_{x,y\boldsymbol{.}x'}$	Partial Kendall correlation between x and y given x'.
ul	upper lip
vl	velum
$\boldsymbol{X}_t$	matrix of acoustic log-energy features
$x(\cdot,\cdot)$	an entry of the matrix $\boldsymbol{X}_t$
ξ_t	set of elementary articulators
$\boldsymbol{y}_t$	EMA values at time t
$\mathbb{Z}$	the set of natural numbers

Symbol	
W	a particular wavelet space
$\boldsymbol{s}$	speech signal
$W_{\boldsymbol{s}}(\cdot,\cdot)$	continuos wavelet transform of a function $\boldsymbol{s}$
ζ_t	distances describing the form of vocal-tract contour

Acronyms

Acronym	
CLP	Cleft lip and palate
ASR	Automatic speech recognition
MFCC	Mel-frequency cepstral coefficients
LSF	Line spectral frequencies
PLP	Perceptual linear prediction
MLP	Multilayer perceptron
RMS	Root mean square error
EMA	Electromagnetic articulography
GMM	Gaussian mixture models
HMM	Hidden Markov models
FFT	Fast Fourier transform
IPA	International phonetic alphabet
VT	Vocal tract
VTR	Vocal tract resonances
LPC	Linear predictive coding
F_1	First formant
F_2	Second formant
F_3	Third formant
CWT	Continuos wavelet transform
DWT	Discrete wavelet transform
MRA	Multiresolution analysis
WPT	Wavelet packet transform
TF	Time-frequency
DFT	Discrete Fourier transform

Acronym	
EM	Expectation maximization
VTLN	Vocal tract length normalization
IA	Inversion approach
EMo	Equation of motion
EC	Equation of continuity
EV	Equation of wall vibration
JND	Just noticeable difference

1 Introduction

The goal of the acoustic-to-articulatory inversion is to estimate articulators movement from the acoustic information contained in the speech signal [94, 105]. g stands for the function that maps the articulatory parameters to the acoustic waveform,

$$\boldsymbol{v} = \mathrm{g}(\boldsymbol{y}) \tag{1-1}$$

where, $\boldsymbol{y}$ is the vector representing articulatory information and $\boldsymbol{v}$ is the acoustic data. Humans can utter very similar sounds using different articulatory configurations; therefore, different combinations of articulatory parameters might be associated to very similar acoustic spectra [8]; that is, $\mathrm{g} : \boldsymbol{y} \mapsto \boldsymbol{v}$ is a many-to-one mapping. This phenomenon causes the inverse mapping $\mathrm{h} : \boldsymbol{v} \mapsto \boldsymbol{y}$ be a one-to-many mapping, thus the articulatory inversion is an *ill-posed* problem [104, 90, 76]. The one-to-many nature of articulatory inversion is studied in [84, 8, 28].

1.1. Motivation

Even though acoustic-to-articulatory inversion offers new perspectives and interesting applications in the speech processing field, it is still an unsolved problem [82, 79, 77]. An adequate system for recovering the articulatory configurations, from the acoustic speech signal, might be used in several applications:

- *Speech therapy*: Visual aids in articulatory training tasks for hearing or speech impaired people [120].
- *Pronunciation trainning*: Computer guided second language learning programs to show correct and incorrect pronunciation [21, 10].
- *Coding*: Low-bit rate coding algorithms since articulators move relatively slowly [95].

- *Speech recognition*: Articulatory parameters representing co-articulatory related phenomena for complementing input feature sets in speech recognition systems [25, 51].

Although the number of potential applications of articulatory inversion is considerable, as reported in [89, 60], the potential uses that motivates present undertaking are mainly two: speech therapy and speech recognition related applications. They are explained with additional detail as follows,

Speech therapy. Regarding speech therapy, if a visual display of the articulators movement were available, it would be easier for the user to obtain an appreciation of their own articulation. The visual feedback offered by the system would allow the analysis and correction of articulatory movements, specially in those cases where it is difficult to observe the movement of articulators. In particular, the development of an inversion system would boost the advancing of speech therapy related applications at the Signal Processing and Recognition Group [1], for which the present work is the first attempt concerning articulatory inversion. The Signal Processing and Recognition Group , at Universidad Nacional de Colombia, has been working on the detection of hypernasality [15, 98, 99] and other problems associated to cleft palate defects. Cleft lip and palate (CLP) may cause functional limitations even after adequate surgical and non-surgical treatment, hypernasality and speech disorders being some of them [96]. An effective assessment system for detecting place of articulation could be useful for correcting pronunciation in cleft palate children.

Speech recognition. Most of the automatic speech recognition systems (ASR) represent the acoustic signal as a chain of sounds [26]. Describing the speech signal as a set of non-overlapping sounds makes difficult the effective modeling of spontaneous speech [51]. The phoneme is a concept that includes different phonetic elements that are produced and modified depending on the context. For example, alveolar plosive phonemes are typically produced by touching the laminal part of the tongue to the alveolar ridge. However, in American English, for example, the point of articulation of alveolar sounds can be moved forward under the influence of dental phonemes or moved backward under the influence of a posterior vowel [46]. Moreover, speech gestures are planned in a coordinated sequence, being controlled by intrinsic and extrinsic muscles, whose actions are relatively slow and overlapped. This circumstance causes the human speech articulators have limited freedom to move, as well as being interrelated and being ruled by inertia-related phenomena. As a consequence, in the

[1] `http://www.unalmzl.edu.co/gta/signal/`

production of a specified sequence of phonemes, articulators spread their influence outside the phoneme range so that substitution of one phoneme by another alters the neighboring segments [43]. This phenomenon is called *co-articulation*.

On the other hand, the electromagnetic articulograph system [2], which records the movement of speech articulators in the midsagittal plane, has made possible the collecting of databases like MOCHA-TIMIT database [3]. The second motivation of this thesis is to exploit the relatively large amount of human parallel acoustic-articulatory speech data which has become recently available in order to improve the understanding of the speech production phenomenon.

1.2. Review of articulatory inversion

Even though several attempts has been made during more than thirty years, the speech researchers still regard the acoustic-to-articulatory inversion as an open issue [52, 48, 79]. Roughly, inversion methods can be divided into two categories: analysis-by-synthesis approaches and nonlinear regression based approaches.

Several articulatory inversion methods are based on the analysis-by-synthesis approach, which is an optimization closed loop procedure that involves the comparison of the spectrum of synthesized speech to measured speech at consecutive frames, for example [100, 104, 76, 82, 79]. In the analysis-by-synthesis approach, the model parameters are adjusted until the articulatory synthesizer output matches the acoustic target, that is, until the synthesized signal is somehow similar to the observed acoustic signal. Figure **1-1** shows a block diagram of an analysis-by-synthesis system [79].

For vowels, the first three formants are commonly used as acoustic features; but, the use of whole spectrum has been proposed, also. The articulatory codebook in Figure **1-1** is used to save approximate and accepted vocal-tract shapes (articulatory configurations). Then, the approximate articulatory configurations are utilized to initialize the optimizer that refines the searching of articulatory parameters until the simulated sound and the measured acoustic data are similar.

On the other hand, nonlinear regression based approaches require a considerable quantity of parallel acoustic-articulatory data. Fortunately, technologies such as electromagnetic

[2] `http://www.articulograph.de/`
[3] `http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html`

articulography have increased the availability of human articulation measurements during speech; therefore, machine learning based methods can be used for the parameters estimation of the nonlinear function relating acoustical and articulatory phenomena. Examples of methods relying on databases of simultaneously collected acoustics and articulatory data are neural networks [90], Gaussian mixture models [110] [77], hidden Markov models [122].

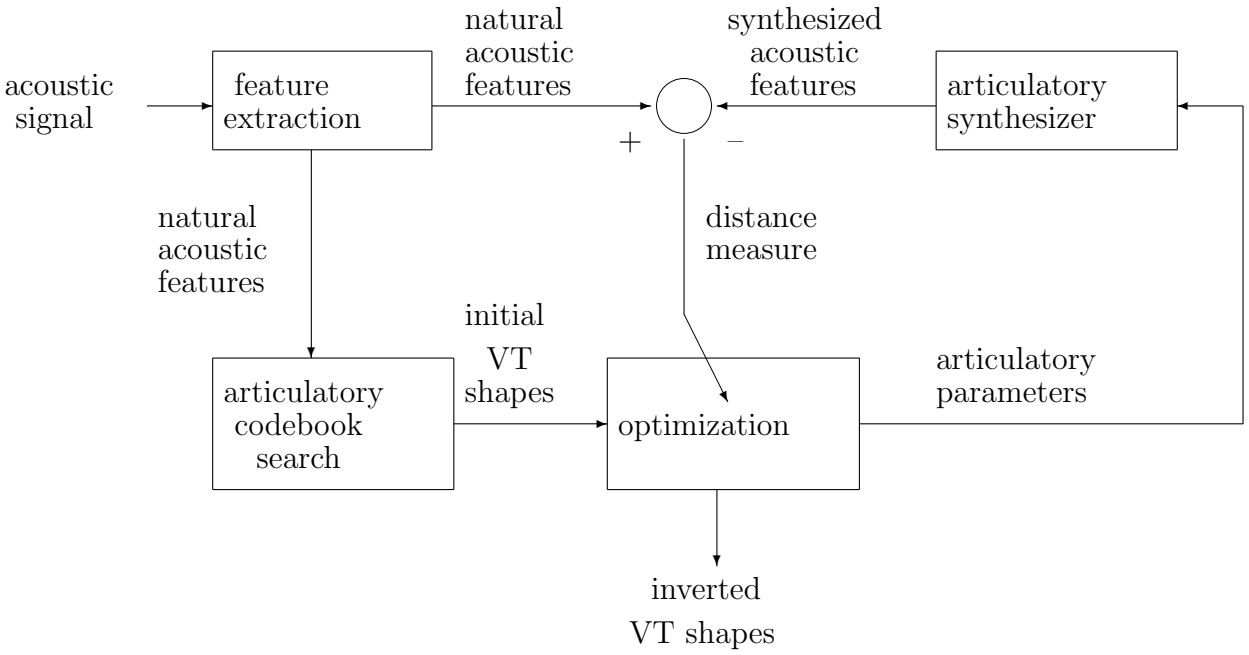


Figure 1-1: Vocal tract inversion system using the analysis-by-synthesis approach. Credits: S. Panchapagesan, see [78].

1.2.1. Measurement of the articulatory phenomenon

Several systems have been developed in order to acquire the articulators movement directly from human subjects: x-ray cineradiography, x-ray microbeam, ultrasound, electromagnetic articulography and magnetic resonance imaging systems. They can be classified from different perspectives, as shown in table **1-4**. Further explanation about each technique is provided:

x-ray cineradiography. It corresponds to the filming of vocal tract during speech production. x-ray images of vocal tract shape have been used in [69, 70]. Although its usefulness,

	x-ray	x-ray microbeam	ultrasound	EMA	MRI
whole VT	yes	no	no	no	yes
tongue imaging	full-length	pellets	full-length	pellets	full-length
velum tracking	yes	yes	no	yes	yes
3D	no	no	no	yes	yes
health hazard	yes	??	no	no	no
nat. articulation	yes	affected	yes	affected	affected
acoustic noise	low	acceptable	acceptable	low	high

Table 1-1: Comparison of vocal-tract shape recording techniques [60].

x-rays are dangerous in case of ration overexposure; therefore, only a very limited set of image sequences are available. Moreover, this technique is not longer used.

x-ray microbeam. The x-ray microbeam system [1] allows for the concurrent observation of lip, jaw, tongue and velum movements by tracking the position of small gold spheres attached to these articulators [116]. In contrast to x-ray cineradiography technique, where the entire head is irradiated, the microbeam generates a small rastered scan of the zone where each pellet is expected to be [4]. The development of x-microbeam system at Wisconsin University did take into account the ensuring of subject safety[5]. Although this technique provides information of only some points of vocal tract, methods to estimate whole tongue shape have been developed. For instance, the authors in [85] propose an algorithm that recovers realistic tongue contours for articulatory databases, based only on the 2D coordinates for the tongue pellets provided in the latter. As a result, realistic tongue shapes can be reconstructed from articulatory databases such as x-ray microbeam (XRMB) databases. Other works involving XRMB data are reported in [80, 105], where acoustic-to-articulatory inversion is performed by using this type of data.

Ultrasound images. Basically, ultrasound systems emit sound waves through the material to be analyzed and it receives the reflected waves resulting from the interaction between the waves and the material. The echoes are sensed and adequately interpreted in order to construct ultrasound images [55]. Ultrasound-based systems are used to tongue motion tracking. They are able to provide the contour of the tongue, however the apex is often

[4] `http://www.biostat.wisc.edu/~myers/ubeam/`
[5] `http://www.psl.wisc.edu/projects/large/microbeam/more-microbeam`

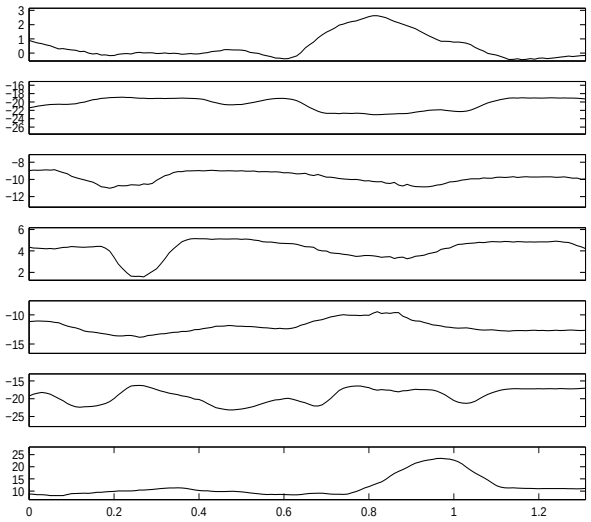


Figure 1-2: **EMA trajectories**.

occluded during speech production by the jaw [34].

Electromagnetic articulography. Electromagnetic articulography (EMA) technology allows the recording of acoustic and articulatory information, but different to x-rays based methods, it allows the collection of a considerable quantity of data. The EMA system is based on the fact that when a spool in introduced in a magnetic field, which varies in a sinusoidal way at a particular rate, a signal with the same frequency is produced in the spool. The provided voltage changes inversely with the distance between the transmitter and the spools, in an approximate way corresponding to the cube to the same distance [89]. Therefore, when measuring the voltages, the distance can be inferred in relation to a particular point of reference. EMA devices are manufactured by Carstens Medizinelektronik GmbH [6]. One of the developed devices is the articulograph AG100, which is capable of determining the Cartesian coordinates x, y of the position of up to 10 sensors in the midsagittal plane (vertical section through the head along the line of symmetry of the face) at a sampling frequency of 500 Hz.

Magnetic resonance imaging. Magnetic resonance imaging (MRI) is a powerful tool that is capable of obtaining 3D images of the vocal tract as well as the tongue shape; however,

[6] `http://www.articulograph.de`

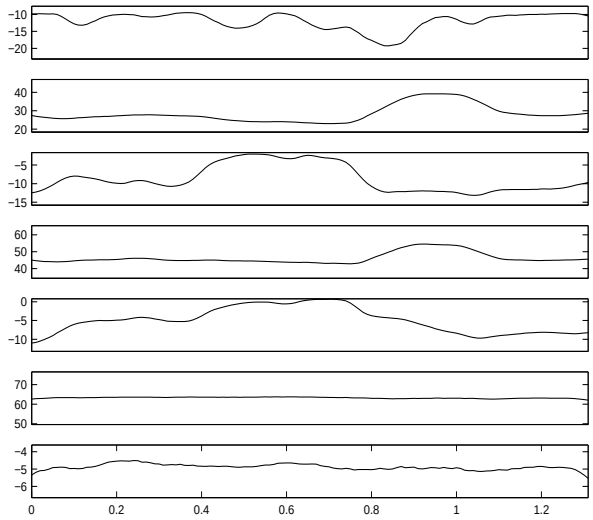


Figure 1-3: **EMA trajectories**.

its use in speech research is restricted due to several reasons [60]:

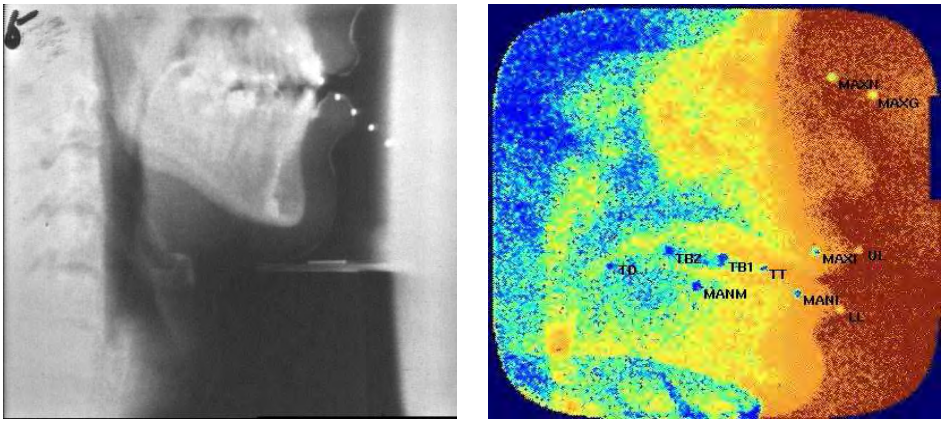
- Low sampling rate, the number of frames per second, restricts the use of MRI to sustained vowels.
- When capturing the MRI images, the speaker needs to be positioned in a painful way, which could cause the articulatory and acoustic patterns be modified.
- The high cost associated with using MRI equipment has restricted its use in speech research.

1.2.2. Inversion methods

Inversion mapping based on training data

Several machine learning based methods have been proposed for solving the articulatory inversion problem, for example codebooks, neural networks, Gaussian mixture models and hidden Markov models.

Codebooks. An articulatory codebook consist of a linked list of vocal tract shape vectors $\boldsymbol{y}$ and related acoustic vectors $\boldsymbol{v}$. Although codebooks are more commonly used in analysis-



(a) (b)

(c) (d)

Figure 1-4: Methods for measuring the articulatory phenomenon. a) x-ray imaging; b) x-ray microbeam; c) ultrasound imaging; d) magnetic resonance imaging (MRI, image of the vocal tract during the production of fricative /ʃ/).
a) source by:`http://psyc.queensu.ca/~munhallk/`
b) source by:`http://www.biostat.wisc.edu/~myers/ubeam/`
c) source by:`http://speech.umaryland.edu/edgetrak.html`
d) source by:`http://www.phon.ox.ac.uk/jcoleman/imaging_links.html`

by-synthesis methods, they have been used as a mean to perform nonlinear regression on real human data in [40]. The dataset is composed of simultaneous articulatory and acoustic measurements of speech corresponding to vowels, vowel-to-vowel transitions, and the consonant /g/. In fact, the speaker produced utterances containing two vowels spoken in a /g/ context with a continuous transition between the vowels. Articulatory measurements are obtained by using an EMA system.

Neural networks. A neural network approach is used in [80] to infer articulation; where, the articulatory information results from tracking movements of gold pellets attached to different articulators in a x-ray microbeam system. Acoustic data is simultaneously recorded and used to infer articulation for the stop consonants /p,b, t, d, k, g/. Differences in behavior are observed between critical and noncritical articulators. Neural networks are also utilized in [90], but using EMA data. The utterances used in this work correspond to sentences that were selected such that the resulting material is phonetically diverse. In addition to MLP, mixture density networks (MDN), neural networks with density functions at the output layer, are used in [90] and their results are compared with the ones obtained by using multilayer perceptrons (MLP) on an acoustic-to-articulatory mapping task. It is observed that the performance for an unseen test set is consistently higher with the MDN than with the MLP. On the other hand, MLP is the articulatory inversion approach used in [**?**] for comparing several acoustic features, different short-time window lengths and different levels of smoothing of the acoustic temporal trajectories in the front-end system. Finally, in the study reported in [52], linear regression and neural network regression are the inversion strategies selected for testing the importance of visual cues for inversion.

Gaussian mixture models. Gaussian mixture models (GMM) with minimum mean-square error (MMSE) and maximum likelihood estimation (MLE) criteria are applied in [110] to determine the trajectory of EMA pellets attached to articulators. They use appropriate static and dynamic features in order to reduce the presence of unnatural movements in the estimated trajectory. It is found that the MLE-based mapping with dynamic features can significantly improve the performance compared with the MMSE-based criterion. Results achieved in [110], in case of using static features, are close to those obtained in [90]. GMMs are also used in [77], but in conjunction with Kalman smoothing. At last, GMMs are used in [7] to analyze the relation between acoustic and articulatory gestures; where, 2D-DCT representations of the the acoustic signal as well as the articulatory signal are used.

Support vector regression. Support vector regression (SVR) technique is tested in [112, 97]. Results in [112] are rather similar to those exposed in previous works [90, 110]. Training process of SVR systems is time consuming in case of large number of patterns as well as it is the finding of hyperparameters. In consequence, the authors in [112] utilized a reduced practical training set in order to decrease training times. This set is constructed by selecting the first out of every five consecutive candidate training examples from a set of patterns estimated at 5 ms rate.

Local regression. Acoustic-to-articulatory inversion using local non-parametric regression and local linear regression has been performed in [6]. Experiments, which are based on EMA data, show no improvement compared with previous methods based on Gaussian mixture models [110] and multiple density networks [88].

Hidden Markov model. In order to take into account the dynamic nature of speech production mechanism, a hidden Markov model (HMM) based inversion method is developed in [38]. The articulatory parameter vector sequence is estimated from the sequence of acoustic parameter vectors and the HMM state sequence by maximizing the a posteriori probability (MAP). Articulatory movements are estimated with an average RMS error of 1.5 mm when using the speech acoustics and the phonemic information, and 1.73 when using only the speech acoustics. The authors in [65] introduce an approach to predict articulatory movements from text. Prediction performance is used to compare three input configurations: a) text input alone; b) audio input alone; and c) both text and audio input together. It is observed that when both the text and the acoustic features are combined, the achieved performance is the highest.

Subject independent approaches. All above cited approaches are subject dependent; thus, they may be unsuccessful if acoustic-articulatory information belonging to the test subject is not included in the training data. In order to overcome this problem, multi-speaker acoustic-to-articulatory inversion based on hidden Markov models (HMM) had been developed in [39, 121]. However, they make use of a stream with information about the phonemes present in the speech signal. A different approach for subject-independent acoustic-to-articulatory inversion approach is proposed in [32]; where, the input acoustic features are transformed into another space such that issues related to inter-subject speaker variability are alleviated. The input space is further partitioned into clusters and then a probability density function is estimated for each cluster. When the probability of generating two acoustic features by

the same cluster is higher compared to other clusters, those feature vectors are assumed to be acoustically close.

Inversion by synthesis

Shirai and Kobayashi in [100] introduced an inversion approach based on nonlinear optimization of articulatory parameters. The process finds those parameters governing the articulatory model such that its output matches the recorded speech signal after glottal and radiation characteristics are removed. The estimation of articulatory parameters is achieved so as to minimize the following cost function:

$$J(\boldsymbol{y}) = (\boldsymbol{v} - g(\boldsymbol{y}))^{\intercal} M (\boldsymbol{v} - g(\boldsymbol{y})) + \boldsymbol{y}^{\intercal} Q \boldsymbol{y} + (\boldsymbol{y} - \boldsymbol{y}_0)^{\intercal} R (\boldsymbol{y} - \boldsymbol{y}_0) \tag{1-2}$$

where, $(\boldsymbol{v} - g(\boldsymbol{y}))^{\intercal} M (\boldsymbol{v} - g(\boldsymbol{y}))$ represents the spectral distance between the output model and the speech wave. Q and R are positive definitive matrices for suitable training, and $\boldsymbol{y}_0$ is the estimation of the previous frame.

A formant-to-articulatory inversion strategy has been studied in [105]. The criterion of optimality involves the minimization of muscle work; where, a second order differential equation is used to describe the relation between muscle work and displacement of articulators. The tissue elasticity coefficient, which is required within the differential equation, is determined experimentally for each articulator. It is also established that first three formants are adequate to compare synthesized and measured speech. Formants are compared in a logarithmic scale in order to resemble somehow the perceptual mechanism. Results are validated in x-ray microbeam data. A similar strategy is used in [103] to perform inversion on unvoiced fricatives. The whole spectrum is utilized for speech representation, instead of only formants. In both cases, the initialization process of the optimization algorithm is made manually.

Codebooks are commonly used in inversion-by-synthesis methods to find the starting points for further optimization process. The ambiguity caused by the one-to-many nature of the inverse problem cannot be resolved by the codebook technique alone because of the necessity to make a choice among multiple articulatory vectors corresponding to the same acoustical vector [104]. However, constraints applied to the dynamics of the articulatory parameters can be used to reduce the uncertainty of the solution. In order to overcome this problem, most studies look for smooth articulatory trajectories under the constraint of matching a given sequence of speech spectra [60]. The elements contained in the codebook

are used mainly as the starting point for further optimization process; where, such elements should adequately span the articulatory space as well as the acoustic space [94]. Recent analysis-by-synthesis methods use codebooks and perform dynamic articulatory inversion by means of following three steps [60]:

- for each acoustic vector representing the current speech frame, a number of articulatory cognates are retrieved from codebook.
- techniques such as dynamic programming are used to obtain an initial articulatory trajectory, which will be the starting point for further optimization processes.
- the articulatory trajectory is optimized such that better acoustic accuracy is obtained.

For the sake of obtaining a codebook that adequately spans the acoustic-articulatory spaces containing a reduced number of elements, Ouni and Laprie in [76] developed a method to construct hyper-cubic tables (multidimensional codebooks) defining articulatory/acoustic relation that guarantees that the acoustical resolution is almost independent of the region in the articulatory space under consideration. They used a strategy of adaptive sampling and local linearity of the relation to decrease the size of codebooks by regrouping samples that span within the linearity domain of an articulatory vector.

Optimization processes involve the use of constraints to reduce the uncertainty of articulatory parameter values. Sorokin [104] presents seven possible kinds of constraints: limitations in the contractive force of muscles involved in the speech production process, which determines the maximum velocity and acceleration of the articulators; anatomy of the vocal tract, which provokes the articulatory parameters to be confined to a certain range of values they cannot overpass; mutual dependence between the articulatory parameters; interdependence between transversal and midsagittal dimensions of the vocal tract; aerodynamic constraints related to the geometry of the vocal-tract when producing different kinds of sound, that is, articulatory configurations provoking turbulent noise are rejected; level of discrepancy between synthesized and measured acoustic signals for the different styles and rates of speech; and finally, constraints regarding the complexity of planning and programming motor commands.

Aerodynamic constraints, limitations on the range of articulatory parameter values and criterion for minimal of muscle work are used in [105]. The constrains used in [104] are: 1) the constraints for the range of articulatory parameters; 2) the constraints for transversal geometric parameters of the vocal tract obtained from the articulatory synthesizer; (3) the requirement of the vocal tract cross-section area minimality, which should prevent the tur-

bulent noise for vowels. On the other hand, Potard and et. al. in [82] incorporate phonetic constraints as well as constraints about the dynamics of articulatory parameters. Results show that these phonetic constraints favor vocal tract shapes close to those realized by the human speaker. The selected inversion strategy is the one developed in [76]. In case of [79], the cost function includes a distance measure between natural and synthesized first three formants, and parameter regularization and continuity terms; where a study of acoustic-to-articulatory inversion for non-nasalized vowel sound through analysis-by-synthesis using Maeda's articulatory model and the XRBM database is performed.

Other constraints proposed by Sorokin [104] dealing with contractive forces, interdependencies between articulatory parameters, or the complexity of the articulatory parameters are too complex to be exploited because there are almost no data available [82].

Finally, as stated in [79], inversion still faces several challenges: 1) complexity of the articulatory-to-acoustic simulation, 2) the one-to-many nature of inverse mapping, 3) incomplete knowledge about the shape and dynamics of the vocal tract for a given speaker. In [79] insufficient training data is mentioned as another challenge; however, there are some recent initiatives tending to overcome this problem. For instance, the authors in [106] introduced a new collection of articulatory speech data from one speaker. It contains volumetric MRI scans of the speaker's vocal tract during sustained production of vowels and consonants, dynamic midsagittal scans of repetitive consonant-vowel production and acoustic recordings of the speech signal.

1.2.3. Inputs to inversion: representation of the speech signal

The type of acoustic input features on regression based methods. Different front-end parameterization methods have been used for acoustic-to-articulatory mapping. Mel-frequency cepstral coefficients (MFCCs) are widely used in automatic speech recognition systems; and, they have been also used in acoustic-to-articulatory mapping systems [20, 122, 110, 6, 31, 121, 28]. Cepstral coefficients are used in [40] to represent the spectrum below 5 kHz. In [90], the input feature set consisted of mel-scale filterbank coefficients. Line spectrum pairs (LSP) are used in [52], as they are closely related to the formant frequencies and the vocal tract shape. Line spectral frequencies (LSF) are used in [65]. However, none of them perform a study to compare the different acoustic parameterizations used in the inversion task.

There is still a scarce quantity of works searching for the best acoustic parameterization

for articulatory inversion. An empirical analysis about the best acoustic parameterization for articulatory inversion was performed in [**?**]; where, the following features were included in the study: Mel-frequency cepstral coefficients, filter banks, line spectral frequencies and perceptual linear prediction (PLP). Multilayer perceptron (MLP) is the selected inversion approach. It was found that best results are generally obtained with acoustic features that are more closely related to the vocal tract like LSF and PLP; however, the achieved improvement is small. On the other hand, formants, which refer to the most relevant resonance frequencies in the vocal tract and hold a close relation with the position of the articulators, were not included in the analysis by [**?**]. The authors in [77] augmented Mel Frequency Cepstral Coefficients with formant trajectories in an system based on GMMs. This action improves the overall performance of the system. The average RMS error reduction is about 3,4 %, and correlation improves by 2,7 %. However, experiments in [77] are not conclusive; therefore, further experiments that reveal the relationship between the formants and the movement of the articulators are to be further provided.

The filterbank structure is optimized in [30] in such a way features provide the maximal information about articulation. The experiments showed that the cochlear filterbank characteristics have an optimal relationship in an information theoretic sense between cochlear filterbank characteristics and acoustic data; that is, speech gestures and the auditory system are well matched to one another.

Number and localization of features in the TF plane Another aspect of front-end parameterization system configuration is the size and delay of the context-window of analysis. Speech gestures are planned movements in a coordinated sequence, being controlled by intrinsic and extrinsic muscles, whose actions are relatively slow and overlapping. This circumstance causes the human speech articulators (jaw, tongue, lips, etc.) to have limited freedom of movement and to be interrelated and ruled by inertia [43]. That is, the information about a phoneme is not localized on that phoneme's region only, but is spread over a substantial segment of the speech signal. Recent experiments support this affirmation, specifically [117, 37] discusses the use of the mutual information applied to estimation of the distribution of the phonetic information in frequency as well as in time. Thus, it is expected that using an adequate context-window around current time of analysis would include more acoustic information related to articulatory movements and therefore it could improve the performance of acoustic-articulatory inversion systems.

On the other hand, the distribution of the articulatory information on the acoustic

speech signal is also important. The question of how the articulatory information, which come from Electro-Magnetic Articulograph (EMA) systems in present work, is coded in the speech signal remains of practical and theoretical relevance. In particular, the knowledge of the distribution of the articulatory influence on the acoustic speech signal is useful in those applications involving articulatory inversion tasks [94, 104].

Several works have pointed out measured differences of performance of acoustic-to-articulatory mapping systems when using different either context–window sizes or positions. Namely, in [80] a neural network was used with 400 inputs covering approximately 200 ms of speech information. The context window size is tuned to get the highest mapping accuracy in a system based upon Gaussian mixture models, as carried out in [110], where the number of input acoustic frames ranges from 1 to 21 (that is, ranging from 10 ms to 200 ms the context-window width). As a result, the greatest mapping accuracy is achieved if properly fixing the number of input acoustic frames (in the specific case, that number turns to be 11). Likewise, there can be found an optimal number of frames for a system based on support vector regression, as reported by [112] where based on a small scale experiment the use of 17 frames (a context–window of about 160 ms of acoustic information) is the optimal choice. Furthermore, it had been established that the error between the estimated articulatory trajectories and the actual ones can be reduced as much as $5\,\% - 10\,\%$ by shifting 14,4 ms in time the window used to predict articulator positions, as suggested in [40]. In [**?**], several time-delays and context–window sizes were used in a system based on neural networks. It was found that the best performance is achieved when the short-time window is centered 15 ms after the articulatory frame; but the improvement is small. Regarding context–window size, it was found the convenient size is 64 ms. Thus, the facts mentioned above suggest the need for a more in-depth study regarding the distribution of the articulatory information immersed in the acoustic speech signal.

The kind of features in analysis-by-synthesis methods Articulatory codebooks are widely used in applications like acoustic-to-articulatory inversion systems. The articulatory-to-acoustic mapping is represented by an articulatory look-up table that associates vectors of articulatory parameters with a parametric representation of the speech signal which is obtained by synthesizing these articulatory parameters through an articulatory model [82].

Commonly, the acoustic representation used is formant-based [76]. Some reasons explain this choice: 1) formant frequencies are phonetically meaningful [61], especially in vowels; 2) the formants have a close relationship with the vocal tract shape [79]; 3) they facilitate the

distance $\mathrm{d}(\mathrm{F}, \hat{\mathrm{F}})$	reference
$\arg\max_{i=1,2,3} \mid \ln \mathrm{F}_i - \ln \hat{\mathrm{F}}_i \mid$; $\arg\max_{i=1,2,3} \mid \frac{\hat{\mathrm{F}}_i}{\mathrm{F}_i} - 1 \mid$	[105, 104]
$\sum_{i=1}^{3} (\mathrm{F}_i - \hat{\mathrm{F}}_i)^2$	[76, 82]
$\sum_{i=1}^{3} (\log \mathrm{F}_i - \log \hat{\mathrm{F}}_i)^2$	[79]

Table 1-2: Examples of distance expressions used to compare formants.

use of phonetic constrains [82]; 4) low dimensional formant vectors are capable of characterizing speech acoustics [79]; and, 5) they avoid problems arising from unknown characteristics of the vocal source [104]. These properties helps to infer the relation between speech signals and tract shapes.

For instance, formants have been used to represent the speech signal in [105, 104, 76, 82, 79]. Because of the nature of analysis-by-synthesis methods, a comparison between synthesized formants and measured formants is performed. To this end, several distances have been utilized, see table **1-2**.

However, formant estimates present inaccuracies, thus, distances in the cepstral of MFCC spaces are preferred. In [75], several candidate distance measures in the cepstral domain are evaluated. To this end, an articulatory codebook was formed by pairs of acoustic vectors and vocal tract impulse responses (without the influence of glottal characteristics). In [93], the study was extended to natural speech. However, the authors evaluated only some distance measures for retrieving the appropriate codebook entries. To evaluate the similarity between the retrieved speech signal and the real one, they used the Just Noticeable Difference (JND) measure, inspired by the work of Ghitza and Goldstein [29].

1.3. Problem statement

Using a wider context-window would improve the performance of acoustic-to-articulatory inversion systems. However, including many frames around current time of analysis increases noticeably the number of input features or regressors. In consequence, the complexity of the regression system, which is defined in terms of the number of parameters of the model to be estimated, also increases. Given a particular number patterns or articulatory-acoustic pairs, increasing the number of parameters of the regression system can lead to poor generalization performance, even if the model is correct [91, 19]. In order to reduce the number of input variables, a feature selection process based on measures of statistical association can be

used [36]. An adequate variable selection process may improve the prediction performance, as well as may provide a better understanding of the underlying process that relates acoustic and articulatory phenomena.

The formants refer to the most relevant resonance frequencies in the vocal tract. The form of the vocal tract changes in the way the articulators move, and in consequence, the resonance frequencies change as well. Consequently, there is a close relation between the position of the articulators and the resonance frequencies. Therefore, including input features well related to the vocal tract mechanism, such as formants, could improve the performance of regression-based inversion systems.

In case of analysis-by-synthesis approaches, formant extraction from a real speech signal (an essential step to testing inversion) is a difficult task and the determination of formants presents inaccuracies [63]. On the other hand, cepstral coefficients are good candidates as an acoustic speech signal representation. They offer the option of separating source and tract characteristics [41] and, from a computational perspective, cepstral coefficients based on the Discrete Fourier Transform are directly calculated from the speech signal and thus, do not introduce any error in the inversion process [78].

The development of distances in the cepstral domain with the performance of distances in the formant domain is of practical importance. A good distance measure between the acoustic speech frame, which is synthesized via the articulatory model, and the acoustic input frame, is required in applications that need to make comparison between synthesized speech frames with real speech frames [75]. This distance measure needs to be insensitive to spectral tilt, as well as formant bandwidths, both of which are associated to glottal variability. In the case of cepstral parameters, this is equivalent to selecting a cepstral distance with the property that the entry with the minimal distance from the given frame in the cepstral domain, is always the same as the entry with the minimal distance in the formant domain.

The previous statement does not hold in general when simple Euclidean distances are considered. Figure (**1-5**) depicts this problem. It can be observed that selecting those codebook entries, having the minimal cepstral distance values, does not lead to the minimal formant distances.

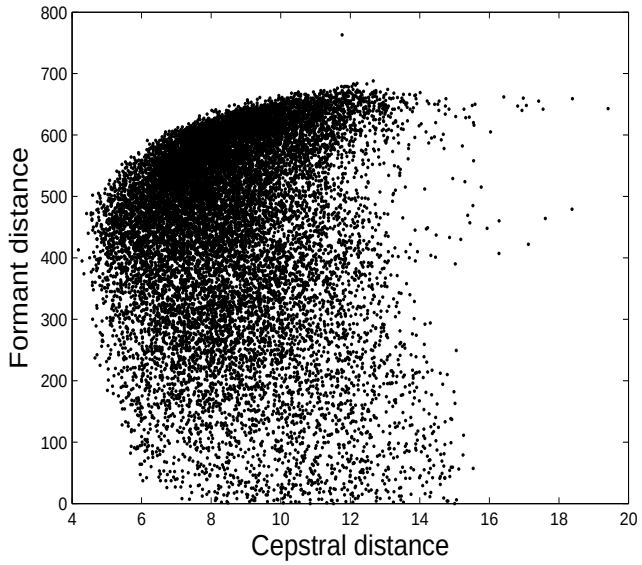


Figure 1-5: First formant vs. (Euclidean) cepstral distances from all entries in the articulatory-to-acoustic codebook to a real speech segment coresponding to an /a/ sound. Minimal cepstral distance does not imply minimal formant distance.

1.4. Thesis outline

It is proposed in present work to use acoustic input features being closely related to position of critical articulators, from the statistical perspective. The proposed method is used to estimate the shape of critical articulator trajectories over fricatives in a speaker–independent way. It requires acoustic-articulatory training data from only one speaker and uses the obtained model to perform articulatory inversion on any arbitrary speaker. Proposed input features are tested in an acoustic-to-articulatory inversion system based on GMM. As a result, the proposed algorithm predicts the direction of movement of the articulators.

Regarding formants, the relation between the articulatory pathways and the formants is analyzed under a statistical perspective. From the analytical point of view we use the Maeda's synthesizer. From the statistical point of view, we estimate statistical association values between the movement of articulators and the resonance frequencies of the vocal tract by estimating mutual information. Then, there is an evaluation of the performance of an articulatory inversion system by adding the resonance frequencies of the vocal tract as entries. This is done in order to show the usefulness of the exposed phenomenon in an articulatory inversion system based on neuronal networks. The statistical analysis is based on real articulatory data which come from an Electromagnetic Articulograph (EMA). The

development of this device is quite recent and it allows us to make measures of speech mechanical activity. Additionally, because of its working principle, it makes possible the acquisition of relatively large quantities of real articulatory data. Therefore, it allows the analysis of the statistical relations between the articulatory and acoustic phenomena in a more reliable way. In case of analysis-by-synthesis methods, the goal of the present work is to find an optimal distance metric for accessing articulatory codebooks. The method is developed for real speech signals, instead of synthesized ones.

The dissertation is organized as follows.

Chapter 2 is a compendium of concepts used through the work. It includes a short introduction to the phonetics of consonants and the kind of features used for the representation of the speech signal.

Chapter 3 presents a method to estimate those acoustic features statistically related to articulators movement. The input feature set is based on time-frequency representation calculated from the speech signal, whose parametrization is achieved using the wavelet-packet transform. The main focus is on measuring the relevant acoustic information, in terms of statistical association, for the inference of articulator positions. Same chapter introduces an approach to finding a cepstral distance minimally affected by glottal variability and closely related to perceptual advantages offered by formants. The process is carried out by optimizing a cost function based on the perceptual distance between real speech frames and synthesized frames.

Chapter 4 describes the data used for the experiments and the their configuration. The following experiments are included: a) experiments for testing cepstral distance measures; b) experiments for testing the contribution of formants on acoustic-to-articulatory mapping systems; and, c) experiments for testing the usefulness relevant maps of TF features.

Chapter 5 shows the results regarding the role of formants in articulatory inversion systems. The resulting cepstral distance, which would replace formants in analysis-synthesis methods, is described. It is also shown the importance of formants as inputs to inversion systems based on nonlinear regression. The analysis is shown from an analytical and a statistical perspective. The former is based on an articulatory synthesizer that simulates the voice signal from the vocal tract. The statistical analysis is based on real data provided by an electromagnetic articulograph.

Chapter 6 exposes the resulting relevant TF features as well as their usefulness. The rank correlation Kendall coefficient is used as the relevance measure. Attained statistical as-

sociation is validated using the χ^2 information measure. The maps of relevant time-frequency features are calculated for the MOCHA–TIMIT database, where the articulatory information is represented by trajectories of specific positions in the vocal tract. Relevant maps are estimated over the whole speech signal as well as on specific phones, for which a given articulator is known to be critical. The usefulness of the relevant maps is tested in an acoustic-to-articulatory mapping system based on Gaussian mixture models. In addition, it is proposed a subject-independent acoustic-to-articulatory mapping method for estimating the shape of the critical articulators movement of fricative sounds. The proposed approach makes use of the acoustic time-frequency features better related to movement of articulators from the statistical perspective. However, the statistical relation estimates are useful for subject–independent articulatory inversion when applied to critical articulators. Testing of proposed features was conducted in an acoustic-to-articulatory mapping system based on Gaussian mixture models. As a result, the proposed approach is able to achieve the inversion with an accuracy, in terms of correlation, similar to the methods used in the subject-dependent that are commonly found in the most contemporary research.

Chapter 7 summarizes the major results and conclusions of the thesis and suggest future directions for research based in this work.

1.5. Publications

Publications in journals:

A. Sepulveda, R. Capobianco, and G. Castellanos. Estimation of relevant time-frequency features using Kendall coefficient for articulator position inference. Accepted for publication in *Speech Communication*, vol. 55 (1), 2013. `http://dx.doi.org/10.1016/j.specom.2012.06.005`.

A. Sepulveda, D. M. Casas, G. Castellanos. Importancia de las frecuencias de resonancia del tracto vocal en la estimación de posiciones articulatorias, *Revista Ingeniería Biomédica*, ISSN 1909-9762. Volumen 6, number 11, 2012.

Alexander Sepulveda y G. Castellanos, Time-frequency energy features for articulator position inference on stop consonants, en *Ingeniería y Ciencia*, vol 8 (16), 2012.

A. Sepúlveda, E. Delgado-Trejos, G. Castellanos and S. Murillo, Hypernasal speech detection by acoustic analysis of unvoiced plosive consonants, en *Revista Tecnológicas*, No.23, Diciembre-2009, p.223-237, ISSN 0123-7799. `http://issuu.com/ideasweb-/docs/tecnologicas_23`

Publications in conferences:

A. Sepúlveda, R. Capobianco-Guido and G. Castellanos, Inference of Critical Articulator Position for Fricative Consonants, in *InterSpeech*, Septiembre de 2012, Portlad, Oregon. http://interspeech2012.org/accepted-abstract.html?id=1205.

A. Sepúlveda, G. Castellanos and R. Capobianco-Guido, Time-Frequency Relevant Features for Critical Articulators Movement Inference, in *European Signal Processing Conference, EUSIPCO, Agosto de 2012*, Bucharest-Rumania. www.eurasip.org/Proceedings/Eusipco/Eusip

A. Sepulveda, J. D. Arias and G. Castellanos, Acoustic-to-articulatory mapping of tongue position for voiced speech signals, in *Advanced Voice Function Assessment International Worshop, AVFA-2009*, Madrid-España. http://www.byo.ics.upm.es/AVFA/resources/AVFA09-

A. Sepulveda, S. Murillo and G. Castellanos, Acoustic-to-articulatory Mapping of Tongue Positions Using Formants, in *V Seminario Internacional Ingerniería Biomédica, Procesamiento y Análisis de Imágenes*, Bogotá-Colombia, Noviembre, 2009. http://sib-sipaim2009.ur

A. Sepulveda, G. Castellanos-Dominguez, and J. Godino-Llorente. Acoustic analysis of the unvoiced stop consonants for detecting hypernasal speech. In *4th International Symposium on Image/Video Communications over fixed and mobile networks*, *ISIVC-2008*. Bilbao-Spain, July 9-11th. http://oa.upm.es/3402/

submitted publications:

Alexander Sepulveda, Yves Laprie, and German Castellanos-Domínguez. Weighted Cepstral Distance Learning for Accessing Articulatory Codebooks. In peer-review process in *ACM Transactions on Speech and Language Processing*.

2 Parametrization of the speech signal

2.1. The nature of speech

2.1.1. Production of speech signal

The lungs are the energy source, which are filled with air by the expansion of the rib-cage and the lowering of the diaphragm. When speaking, air is forced out of the lungs along the trachea. The velocity at which the air exits the lungs serves to control the volume of the produced speech signal [27]. The first section of the vocal apparatus is the larynx, which is the place where the vocal folds are located. The gap between the vocal folds is called the glottis, see Figure **2-1**.

If the glottis is open, air passes freely through the glottis in order to create *voiceless* sounds; for example the /s/ in study /ˈstʌdi/. In contrast, when the vocal folds are held close together, but not firmly closed, the air builds up behind them until it reaches sufficient pressure to force them to separate, causing the pressure to drop. Later, the folds close and the pressure begins to increase once again. This rapidly opening and closing process makes the air exits the glottis in short bursts. Sounds produced in this way are called *voiced*, for example the vowels. Once the air has passed through the glottis, it is directed into either just the mouth, or the mouth and the nose simultaneously, depending upon the position of the velum [17]. Sounds made with the velum raised are named *oral* sounds; while those produced with the velum lowered (with air passing through both cavities nose and mouth), are referred to as *nasal* sounds.

The shape of the vocal tract varies with time due to motions of the lips, jaw, tongue, and velum. This phenomenon makes the sounds be transformed by the vocal-tract impulse response, whose frequency characteristics depends on the particular arrangement or configuration of the articulators. The resonance frequencies resulting from a particular articulatory configuration helps to characterize the sound corresponding to a given phoneme, specially

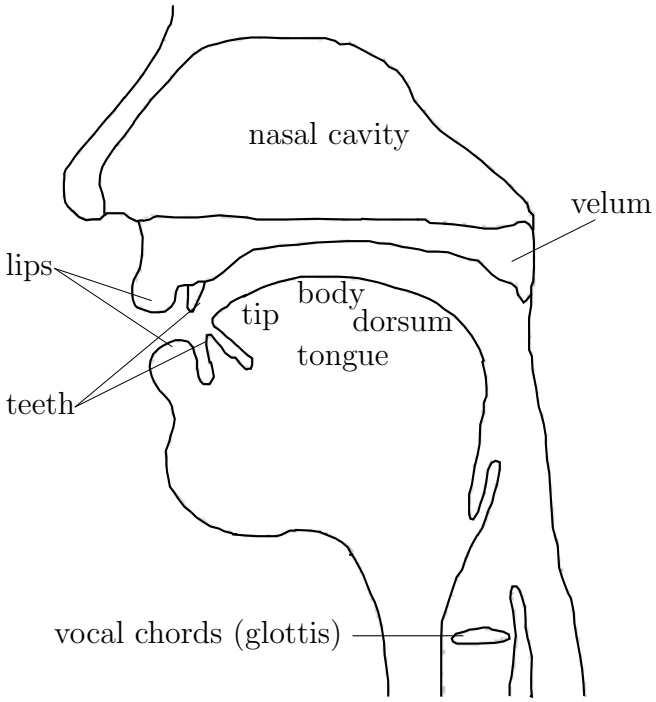


Figure 2-1: Fig:Speech production mechanism

vowels [9]. In vowels, these resonance frequencies are strongly related to formants [17].

But not only vowels make part of the spectrum of speech sounds, consonants also form part of speech sounds. A traditional system for describing phonemes associated to consonants is by using the following categories: manner of articulation, place of articulation and voicing [51]. Table **2-1** shows the classification of English consonants. Manner of articulation deals with airflow in the vocal tract: whether it flows through the oral and/or nasal cavities, and the degree of any major vocal tract constrictions [17].

The categories for manner of articulation are: vowels, glides, liquids, nasal, fricatives and stops (plosives). A plosive consonant is formed by momentarily blocking the oral cavity at some point in order to build up the pressure behind the oral closure, which is then released as a minor explosion [99]. The fricatives are characterized by the formation of a narrow constriction somewhere in the vocal tract followed by the development of turbulent airflow [50]. Glides (also called semivowels) resemble vowels, but they have a very high tongue position, which causes a narrow vocal tract constriction barely wide enough to avoid frication [17].

Place of articulation, which is usually associated to consonants, refers to the location where the vocal tract is the most constricted. Following regions are identified [41, 17]:

- *Bilabials* have their major constriction at the lips.
- In *labiodentals*, the lower lip contacts the upper teeth.
- For the production of *alveolars*, speaker approximates the front part of the tongue, the tip or the blade, to the alveolar ridge.
- *Palatals* have approximation or constriction on or near the roof of the mouth, called the palate.
- In case of *velars*, the tongue dorsum constricts with the region near the velar flap.

	Labial	Labio-dental	Dental	Alveolar	Palatal	Velar	Glottal
Plosive	p b			t d		k g	ʔ
Nasal	m			n		ŋ	
Fricative		f v	θ ð	s z	ʃ ʒ		h
Retroflex				r			
Lateral				l			
Glide	w				j		

Table 2-1: The consonants of English arranged by place (columns) and manner (rows).

2.1.2. Introduction to the acoustics of phonemes

Vowels. Some rules of thumb for relating vowel formant frequencies to tongue configuration have appeared in literature, for example that F_1 (first formant) varies mostly with the tongue height; and, F_2 changes mostly with the tongue advancement. However, there are exceptions to these rules. Roughly, low vowels (/a, ɑ, ʌ/) have a high F_1 value and high vowels (/i, ɪ, u, o, ʊ/) have a low F_1 frequency. Back vowels (/ɔ, ʊ/) have a low F_2 and typically a small $F_2 - F_1$ difference, whereas front vowels (/i, ɪ, ɛ/) have a relatively higher F_2 frequency and large $F_2 - F_1$ difference [50]. The IPA chart corresponding to vowel sounds is shown in figure **2-2**[1].

Stops. Place of articulation in stops can be acoustically cued by using spectral moments and formant transitions [50], as follows,

[1]This image was taken from `http://www.langsci.ucl.ac.uk/ipa/vowels.html`

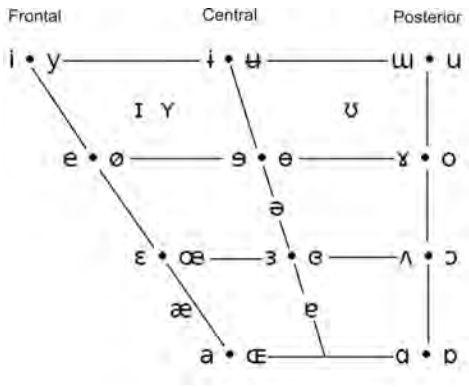


Figure 2-2: Grid for representing the vowels using the IPA system (International Phonetic Alphabet, IPA).

- Bilabials: The spectral moment has relatively low spectral mean, high skewness, and low kurtosis. F_2 frequency increases from stop release into following vowel.
- Alveolar: The spectral moment presents a relatively high spectral mean, low skewness, and low kurtosis. F_2 frequency decreases from stop release into following vowel except for the high front vowels.
- Velar: The spectral moment has a relatively low spectral mean, high skewness, and high kurtosis, probably reflecting compact spectrum. F_2 and F_3 have a wedge-shaped pattern in which they are initially nearly fused but separate in frequency during the transitions.

Fricatives. The role of the second and third formant transitions F_2 and F_3 in detecting the place of articulation in fricatives were investigated by several researchers. The common conclusion is that the spectral shape is a much more important cue for the place of articulation than formant transitions. And that if they played any role, it is for discriminating between labiodentals and dentals [2]. Apart from this, perceptual experiments have shown that formant transitions are not a primary cue in the place of articulation detection [2].

Regarding alveolar fricatives, their spectrum contains relatively higher frequency energy than the spectrum for palatals, as reported in [50]. In case of adult male speakers, the major region of noise energy for the alveolar fricatives lies above 4 kHz. In contrast, the palatal fricatives have significant noise energy extending down to about 3 kHz. These cutoff values are only approximate. In case of voiceless fricatives s and ʃ, the skewness of the spectral moment seems to be an effective distiguishing characteristic. In [107], it is reported that the labiodental and dental fricatives have no major spectral prominence. For the palato-alveolar

fricative ∫ there are spectral peaks localized at 2600 and 2300 Hz, approximately.

Nasals. A nasal consonant is produced by making a complete closure with one of the articulators, while maintainig the velopharyngeal port open [107]; then, the air escapes through the nose is. Several parameters have been used for detecting nasal consonants and nasalized vowels. In particular, the authors in [83] reported the development of a set of acoustic parameters that can be extracted automatically and reliably in a speaker independent way. They reported some observations: first, there is energy at very low frequency and the sudden drop in energy above it for nasals, then, they used the ratio of energies between 0-320 and 320-5360 Hz as an acoustic parameter; second, nasals tend to have lower frequency value for the first spectral prominence as compared to semivowels.

The nasalization of the acoustic signal applies not only to the nasal consonants but also to certain sorrouding sounds, particularly vowels. Hence, the acoustic cues for nasalization often can be found beyond the nasal consonant segment [50].

2.1.3. Critical articulators.

It is shown in [80] that certain articulators play a more significant role to the production of a given phone than other articulators. These articulators are called *critical articulators*. It is reported in [80] that the movement of critical articulators had a greater range but that are less variable, compared to non-critical articulators. In consequence, relatively good correlations and relatively poor RMS errors are obtained for the case of critical articulators. The authors is [24] offer additional evidence regarding the the link between critical articulators and speech acousctics; where, a method that weigths articulatory parameters according their ability to predict a given acoustic sound is utilized. Results show that those articulators that are critical for the especified phomene are weighted greater than non-critical ones. For instance, the phonemes /f m p/ received a large weight for the lower lip, which is considered the critical articulator for bilabial or labiodental phones.

In order to obtain an estimated of critical articulators for phones, the International Phonetic Alphabet (IPA) can be used. IPA descriptors are widely used, have been refined during decades and they are able to represent phones found in human languages [43]. For instance, as shown in table **2-1**, the critical articulator of fricative sounds /f v/ is lower lip; and, in case of /t d/, the critical articulator is the tongue tip.

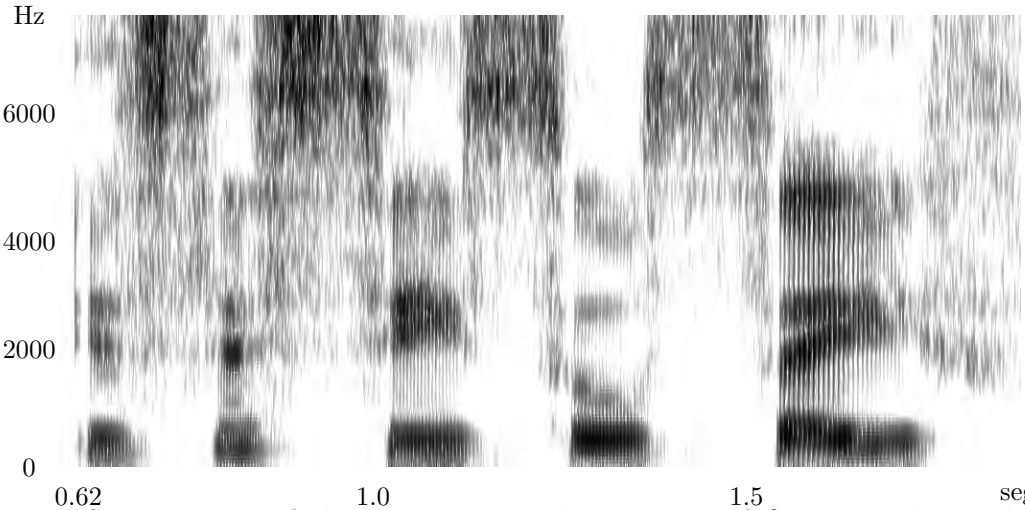


Figure 2-3: Spectrogram of the utterance *Is this seesaw safe?*. It was obtained using the software *WinSnorri* [59].

2.2. Speech signal representation

The speech signal conveys several types of information. The information on a signal can be more easily interpreted in a certain representation than in others; therefore, the choice of representation is of great importance. For example, the information infered by humans from the visualized waveform is scarce and limited; however, a lot of information becomes easily accesible when transforming the speech signal into the spectral domain. Indeed, the reading of spectrograms has been used by experts to infere the phonetic information [123]. The spectrogram for the phrase *Is this seesaw safe?* belonging to speaker fsew0 in MOCHA-TIMIT database is shown in figure **2-3** [2].

Speech signals are transformed to a lower dimensional feature set or feature vectors by using several feature estimation techniques. Even though there are many options to represent acoustic signals, only some of the most commom ones are included in present work.

2.2.1. Cepstrum

The real cepstrum (the inverse transform of the logarithm of the speech power spectrum) is defined by,

$$\boldsymbol{c}_s = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log \mid S(\omega) \mid e^{j\omega t} d\omega \tag{2-1}$$

[2]The spectrogram is estimated by using the software *WinSnorri* `www.loria.fr/~laprie/WinSnoori/`

where $S(\omega)$ is the spectrum of the signal $\boldsymbol{s}(t)$.

Frequently, the natural or the base 10 logarithm is used in this computation, although any base can be used [16]. Commonly, the DFT (discrete Fourier transform) is used for its computation [17]; therefore,

$$\boldsymbol{c} = \frac{1}{N_e} \sum_{k=0}^{N_e-1} \log \mid S(k) \mid e^{j2\pi kt/N_e} \quad \text{for} \quad t = 0, 1, \ldots, N_e - 1$$

where

$$\boldsymbol{S}(k) = \sum_{t=0}^{N_e-1} \boldsymbol{s}(t)\, e^{-j2\pi kt/N_s} \tag{2-2}$$

is the discrete spectral representation of the real signal $\boldsymbol{s}(t)$ and N_e is the number of samples in the signal $\boldsymbol{s}(t)$.

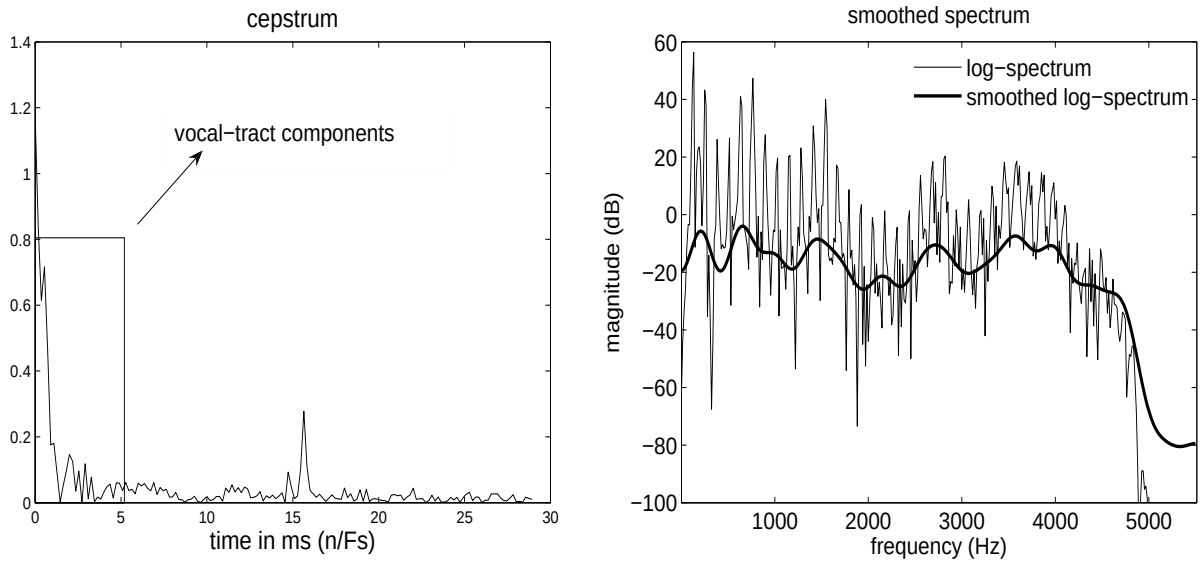
A simple model of the speech process considers the voiced sounds to be produced by quasi-periodic pulses of air caused by vocal cords vibration. Glottal pulses are generated, which excite the vocal tract to finally produce speech [13]. The result of this effect in the time domain can be represented by the convolution function, and in the frequency domain, the output to this phenomenon is represented by the multiplication of the vocal-tract frequency representation with the source representation. The real cepstrum operator transforms the convolution operation into addition.

Let the speech spectrum be $\boldsymbol{S} = \boldsymbol{E} \cdot \boldsymbol{G}$, where $\boldsymbol{E}$ and $\boldsymbol{G}$ represent the excitation and vocal-tract spectra, respectively; then $\log \boldsymbol{S} = \log(\boldsymbol{E} \cdot \boldsymbol{G}) = \log \boldsymbol{E} + \log \boldsymbol{G}$. Since $\boldsymbol{G}$ consists mostly of the spectrum varying slowly with frequency while $\boldsymbol{E}$ is much more active or irregular (owing to the harmonics or noise excitation), contributions due to $\boldsymbol{E}$ and $\boldsymbol{G}$ can be linearly separated [17]. This property makes cepstrum useful to separate information about the vocal tract contributions from the short-time speech spectrum [86].

Since the focus of acoustic-to-articulatory inversion is the inference of the vocal-tract shape, it is better to use a representation that removes the speech source influence. To acomplish this end, cepstral smoothing as well as formants are a good candidates. However, compared to formant parametrization, cepstrum features can be more easily and reliably estimated [78].

An example of the smoothed spectral response of the vocal-tract is shown in Figure **2-4**. The cepstral values are computed for a recording of a male speaker at 11025 Hz. The

values above 5,4 ms are used to estimate the smoothed spectrum of Figure **2-4**(b). The peaks in the smoothed spectrum are related to the formants; but, it is possible the spectrum estimated by ceptral smoothing gives false formants in between the actual ones.



(a) (b)

Figure 2-4: Cepstral coefficients for a speech frame sampled at 11025 Hz. a) Cepstral signal. The range enclosed by the rectangle indicates the first 5,4 ms in the cepstral domain, which contains the coefficients used to reconstruct the smothed spectrum of figure in b). b) Vocal-tract smoothed spectrum response when using the coefficients belonging to the first 5,4 ms.

2.2.2. Formants

Resonance frequencies of the human vocal tract are the natural frequencies, or eigen-frequencies, of the air path in the vocal tract from glottis to lips, and the air path is shaped principally by the tongue, jaw, and other articulators [62]. Since such vocal tract resonances (VTRs) describe the physical system, they are required to exist at all times, even with weak or no measurable emitting acoustic signals. By contrast, formants are defined in the acoustic domain. Formants are associated with peaks or prominences in the smoothed power spectrum of the acoustic signal of speech; that is, they are related to local maxima in the displayed amplitude spectrum that is not due to source-spectrum related properties [66], see Figure **2-5**. Taking into account the acoustic definition, formants would disappear during complete consonantal closure [62]. When anti-resonances are present, as in most consonantal sounds, the acoustic effect of the underlying resonance frequencies are often obscured [63, 17].

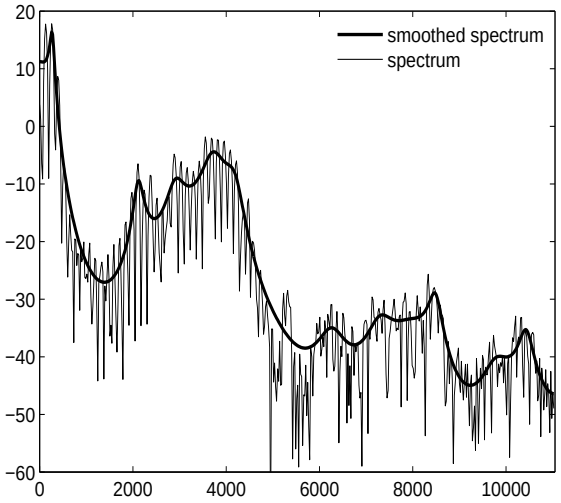


Figure 2-5: Spectrum for the sound /i/ of a male Spanish speaker. Plot in bold stands for the smoothed spectrum by using linear predictive coding. Peaks correspond to formant frequencies.

Most experience with synthetic speech lends support to formant patterns as primary cue for vowel perception. When vowels have been synthesized using formant frequencies estimated from natural speech, the results have been generally satisfactory; for example, the Klatt synthesizer implemented in the software *WinSnoori* [59]. On the other hand, experiments based on articulatory synthesizers showing the relation between formants and vocal tract representations have been developed [64]; where, it was found that F_1 varies with tongue height and F_2 varies with tongue advancement. In addition, some observations are made in [70]: For vowels, like /o/ and /u/ in Spanish, the effect of the jaw appears on the F_1 value, and those of the tongue dorsal position primarily appears on F_2. It is also observed that F_1 and F_2 frequencies are quite stable against the variation in the lip aperture; however, the effects of the rounding can apperar as a lowering of F_3 frequency, resulting in the typical French sound change from /i/ to /y/, whose articulatory configurations are similar except by the shape of lips.

Several algorithms and software packages have been developed for VTRs/formants estimation; for example *WaveSurfer*, which is a popular open source tool [63]. Its performance has been compared with recently developed algorithms [102]. *WaveSurfer* makes use of the algorithm introduced in [109] for automatic $F_1/F_2/F_3$ tracking.

2.2.3. Representation based on wavelet packet transform

It must be highlighted that the acoustic features can be represented by using different known time–frequency approaches. Nonetheless, the main motivation for using wavelet packets is that they can be efficiently implemented with relatively low computational cost [101]. In addition, they offer an alternative for detecting sudden bursts in slowly varying signals [3], which is a phenomenon observed in stop consonants.

The *continuous wavelet transform* (CWT) of a function $\boldsymbol{s}(t)$ is defined as follows,

$$W_s(u,b) = \langle \boldsymbol{s}(t), \psi_{u,b} \rangle = \int_{-\infty}^{\infty} \boldsymbol{s}(t) \frac{1}{\sqrt{b}} \psi(\frac{t-u}{b}) dt \tag{2-3}$$

where each time-frequency atom $\psi_{u,b}$ corresponds to a shifted and scaled version of the mother wavelet ψ. b is the scaling parameter and u is the shifting variable.

In order to obtain a more practical transform, the parameters b, u are constrained to discrete values. In consequence, u is defined as $u = nu_0 b_0^j$, where u_0 is a fixed value and $n \in \mathbb{Z}$ [14]. Thus, discretized wavelets take the form,

$$\psi_{j,n}(t) = b_0^{-j/2} \psi(b_0^{-j} t - nu_0) \tag{2-4}$$

If $b_0 = 2$ and $u_0 = 1$, then it exist a set of functions $\{\psi_{j,n}\}$, where $\psi_{j,n}(t) = 2^{-j/2}\psi(2^{-j}t - n)$, such that the set $\{\psi_{j,n}\}$ is an orthonormal base in $\mathbf{L}^2(\mathbb{R})$.

Discrete wavelet transform

The main idea of *discrete wavelet transform* (DWT) is the process of multiresolution analysis (MRA) proposed by Mallat [35]. DWT is inherently tied to the concept of multiresolution analysis. A MRA consists of a sequence $\{V_j\}_{j\in\mathbb{Z}}$ of closed subspaces of $\mathbf{L}^2(\mathbb{R})$ which have the following properties [73, 4]:

- *Containment:*
  $\forall j \in \mathbb{Z}$, $V_{j+1} \subset V_j$; that is $\cdots V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \cdots \subset \mathbf{L}^2(\mathbb{R})$.
  V_2, V_1 stands for the coarser spaces while V_{-1}, V_{-2} correspond to the finer spaces.
- *Completeness:*
  $\bigcup_{j\in\mathbb{Z}} V_j = \mathbf{L}^2(\mathbb{R})$; $\bigcap_{j\in\mathbb{Z}} V_j = \{\emptyset\}$
- *Scaling:*

$\forall j \in \mathbb{Z};\ \boldsymbol{s}(t) \in V_j \Longleftrightarrow \boldsymbol{s}(2t) \in V_{j-1}$

- *The basis:* There is a function $\phi(t) \in V_0$ such that the set $\{\phi_{j,n} = \phi(t-n); n \in \mathbb{Z}\}$ is an orthonormal basis for V_0.

The projection of $\boldsymbol{s}$ on V_j, denoted $P_{V_j}\boldsymbol{s}$, is obtained by using the set of orthonormal basis $\{\phi_{j,n}\}$,

$$P_{V_j}\boldsymbol{s} = \sum_j \boldsymbol{a}_j(n)\phi_{j,n} \tag{2-5}$$

where $\boldsymbol{a}_j(n) = \langle \boldsymbol{s}, \phi_{j,n} \rangle$.

The approximations of $\boldsymbol{s}$ at the scales 2^j and 2^{j-1} are respectively equal to their orthogonal projections on V_j and V_{j-1}. In addition, $V_0 = \text{span}\{\phi(t-n)\}$ and $V_{-1} = \text{span}\{\phi(2t-n)\}$. Let W_j be the orthogonal complement of V_j in V_{j-1}; that is, $V_{j-1} = V_j \oplus W_j$. Thus, the approximation $P_{V_{j-1}}$ can be expressed as a sum of projections of s onto V_j and W_j as follows,

$$P_{V_{j-1}}\boldsymbol{s} = P_{V_j}\boldsymbol{s} + P_{W_j}\boldsymbol{s} \tag{2-6}$$

$P_{V_j}\boldsymbol{s}$ is the low-frequency part of $\boldsymbol{s}$ on V_j while $P_{W_j}f$ is the high-frequency component; where, $P_{W_j}\boldsymbol{s}$ can be obtained by using the expansion,

$$P_{W_j}\boldsymbol{s} = \sum_n \langle \boldsymbol{s}, \psi_{j,n} \rangle \psi_{j,n} \tag{2-7}$$

Let be the function $\psi(t) \in W_0$, such that $W_0 = \text{span}\{\psi(t-n)\}$; therefore, this function $\psi(t)$ is the wavelet function associated with the multiscale analysis. $W_j = \text{span}\{\psi(2^{-j}t-n)\}$, where, $\{\psi_{j,n}(t)\}$ is also a set of orthogonal functions.

Given that $V_{-1} = \text{span}\{\phi(2t-n)\}$ and $\phi(t) \in V_0$ implies $\phi(2t) \in V_{-1}$, $\phi(t)$ can be expressed in terms of $\phi(2t)$, as follows,

$$\phi(t) = 2\sum_n h_0(n)\phi(2t-n) \tag{2-8}$$

where the coefficient set $\{h_0(n)\}$ are the interscale basis coefficients, which is also a low-pass filter. Similarly, the band-pass wavelet function can be expressed as a linear combination of

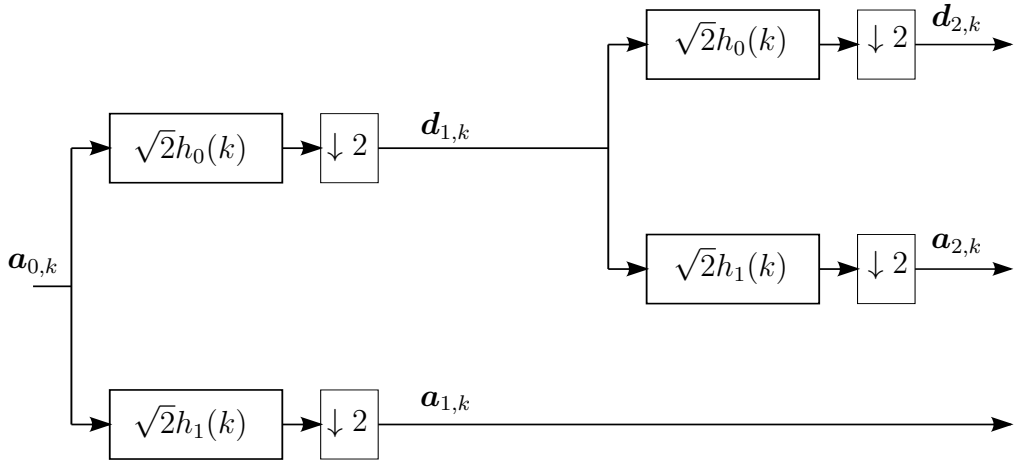


Figure 2-6: Multiresolution decomposition structure

$\phi(2t)$:

$$\psi(t) = 2\sum_n h_1(n)\phi(2t-n) \tag{2-9}$$

If assuming the function $\boldsymbol{s} \in V_0$. Then, $\boldsymbol{s}$ can be represented in terms of $\phi(t-n)$,

$$\boldsymbol{s}(t) = \sum_n \boldsymbol{a}_{0,n}\phi(t-n) \tag{2-10}$$

where $\boldsymbol{a}_{0,n} = \langle \boldsymbol{s}, \phi(t-n)\rangle$. Next, since $V_0 = V_1 \oplus W_1$, $\boldsymbol{s}$ can be expressed as the sum of the two functions $\phi_{1,n}(t)$ and $\psi_{1,n}(t)$,

$$\begin{aligned} \boldsymbol{s}(t) &= P_{V_1}\boldsymbol{s} + P_{W_1}\boldsymbol{s} \\ &= \sum_n \boldsymbol{a}_{1,n}\phi_{1,n}(t) + \sum_n \boldsymbol{d}_{1,n}\psi_{1,n}(t) \end{aligned} \tag{2-11}$$

It is shown in [4] that $\boldsymbol{a}_{1,n}$ and $\boldsymbol{d}_{1,n}$ can be obtained from $\boldsymbol{a}_{0,n}$ by using the expressions,

$$\begin{aligned} \boldsymbol{a}_{1,n} &= \sqrt{2}\sum h_0(k-2n)\boldsymbol{a}_{0,k} \\ \boldsymbol{d}_{1,n} &= \sqrt{2}\sum h_1(k-2n)\boldsymbol{a}_{0,k} \end{aligned} \tag{2-12}$$

The next step of decomposition is also performed; thus, $V_1 = V_2 \oplus W_2$. Similarly, the coefficients for the projections $P_{V_2}f$ and $P_{W_2}f$ are obtained as follows,

$$\begin{aligned} \boldsymbol{a}_{2,n} &= \sqrt{2}\sum h_0(k-2n)\boldsymbol{a}_{1,k} \\ \boldsymbol{d}_{2,n} &= \sqrt{2}\sum h_1(k-2n)\boldsymbol{a}_{1,k} \end{aligned} \tag{2-13}$$

The relations just described are depicted in Figure **2-6**. The discrete wavelet transform of order j is obtained by carrying out the process shown in following equation,

$$\begin{aligned} \boldsymbol{a}_{j+1,n} &= \sqrt{2}\sum h_0(k-2n)\boldsymbol{a}_{j,k} \\ \boldsymbol{d}_{j+1,n} &= \sqrt{2}\sum h_1(k-2n)\boldsymbol{a}_{j,k} \end{aligned} \tag{2-14}$$

Wavelet packet transform

The wavelet packet transform (WPT) is a generalization of the discrete wavelet transform. In case of DWT, the approximation coefficients are used to perform additional decompositions. On the other hand, WPT uses same MRA process as in DWT, but the process is also applied to detail coefficients. In consequence, a tree-like structure is generated.

A space V_j can be decomposed into the two subspaces V_{j+1} and W_{j+1} by using the set of orthogonal basis,

$$\{\varphi_{j+1}(t-2^{j+1}n)\}_{n\in\mathbb{Z}} \quad \in \quad V_{j+1} \quad \text{and} \quad \{\psi_{j+1}(t-2^{j+1}n)\}_{n\in\mathbb{Z}} \quad \in \quad W_{j+1}$$

For computing WPT, detail spaces W_j are also decomposed; which generates a binary tree. Any node of a binary tree is labeled by its depth j and the number of q nodes that are on its left at the depth j, see Figure **2-7**. Each node (j,q) corresponds to a space W_j^q. The root of the tree is $W_0^0 = V_0$. Let be the space $W_j^q = \text{span}\{\psi_j^q(t-2^j n)\}_{n\in\mathbb{Z}}$. The functions forming the bases at the two children nodes are defined by,

$$\psi_{j+1}^{2q}(t) = \sum_{n=-\infty}^{\infty} h(n)\psi_j^p(t-2^j n) \tag{2-15}$$

and

$$\psi_{j+1}^{2q+1}(t) = \sum_{n=-\infty}^{\infty} g(n)\psi_j^q(t - 2^j n) \tag{2-16}$$

In [73] it is shown that $\mathcal{B}_{j+1}^{2q} = \{\psi_{j+1}^{2q}(t - 2^{j+1}n)\}_{n\in\mathbb{Z}}$ and $\mathcal{B}_{j+1}^{2q+1} = \{\psi_{j+1}^{2q+1}(t - 2^{j+1}n)\}_{n\in\mathbb{Z}}$ are orthonormal bases of the two orthogonal spaces W_{j+1}^{2q} and W_{j+1}^{2q+1} such that,

$$W_{j+1}^{2q} \oplus W_{j+1}^{2q+1} = W_j^q \tag{2-17}$$

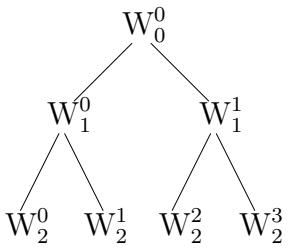


Figure 2-7: Nomenclature in a typical wavelet packet tree.

At each decomposition level, the WPT splits the TF plane on rectangles of constant area. TF atoms became wider in time and narrower in frequency as the decomposition level is increased. Since each level generates a particular partition of the TF plane, a family of TF splits are obtained. Therefore, the options for representing the signal s get noticeable increased.

Mel-like WP representation

Frequency splitting of the time–frequency plane can be generated with the WPT having frequency bands spacing similar to the Mel scale, as proposed in [23]. In the first place, a full three level WPT decomposition is performed, which splits the frequency components within the range $[0, 8]$ kHz into eight bands; where each is of 1 kHz bandwidth approximately. Then, energy localized in the bands $[4, 5]$ kHz, $[5, 6]$ kHz, $[6, 7]$ kHz, and $[7, 8]$ kHz produce the coefficients 21^{st}, 22^{nd}, 23^{rd}, and 24^{th}, respectively. The band within $[3, 4]$ kHz is decomposed once to achieve a couple of bands, ($[3, 3{,}5]$ kHz and $[3{,}5, 4]$ kHz), that generate the 19^{th} and 20^{th} filter banks. Next, the $[2, 3]$ kHz band is selected and split out into 4 bands of 250 Hz bandwidth each. The frequency band of $1 - 2$ kHz is further decomposed applying two level WPT decomposition, thus resulting in four 250 Hz subbands. The frequency bands of $[1, 1{,}25]$ kHz and $[1{,}25, 1{,}5]$ kHz are once more further decomposed, thus increasing the number of

bands to six in the $[1, 2]$ kHz range. Finally, the lowest band of $[0, 1]$ kHz is decomposed by applying a full three level WPT decomposition, and therefore, dividing the $[0, 1]$ kHz band into eight subbands (1^{st} to 8^{th} filter banks), where each one is 125 Hz bandwidth, approximately. Likewise, to accomplish the time plane partition, the acoustic speech signal is parameterized using 20 ms frames and $\Delta t = 10$ ms steps, so a rate frame of 100 Hz is performed [90]. Acoustic information within time interval ranging from $t - t_a = t - 200$ ms to $t + t_b = t + 300$ ms is parameterized.

As a result, the time–frequency information is represented by the scalar valued logarithmic energy features $x(t + d, f_k) \in \mathbb{R}$, where the set $\{f_k : k = 1, \ldots, N_f\}$ appraises the $N_f = 24$ frequency components, where $d \in [t_a, t_b]$ is the time–shift variable. A resulting matrix of acoustic log–energy features $\boldsymbol{X}_t \in \mathbb{R}^{N_t \times N_f}$ (with $N_t = (t_b - t_a)/10$ ms) is attained for each window analysis at the time position t of the articulatory configuration $\boldsymbol{y}_t = \{y^m(t) : m = 1, \ldots, N_c\} \in \mathbb{R}^{N_c \times 1}$, where m denotes the m–th channel and $N_c = 14$ is the number of EMA channels. So, column vector $\boldsymbol{x}_{t+d} = \{x(t + d, f_k)\} \in \mathbb{R}^{N_f \times 1}$, of TF matrix $\boldsymbol{X}_t$ comprises the set of N_f energy features estimated as follows [23]:

- Computation of WPT of the speech frame at time $t+d$, by using Daubechies compactly supported wavelets with six vanishing moments, as in [23].
- Calculation of the energy of each frequency band that results from the sum of square values of the coefficients contained in the WPT–related nodes W_l^p (Table **2-2** shows the W_l^p WPT–nodes related to each filter bank f_k). Then, logarithmic operation is performed over attained set of energy values.

2.2.4. Filter-banks

Given the DFT of an input signal, as provided in 2-2; a filterbank of M triangular filters is defined as follows [41],

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{2(k - f(m-1))}{(f(m+1) - f(m-1))(f(m) - f(m-1))} & f(m-1) < k < f(m) \\ \frac{2(f(m+1) - k)}{(f(m+1) - f(m-1))(f(m+1) - f(m))} & f(m) < k < f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{2-18}$$

The $f(m)$ are provided by the expression,

$$f(m) = \frac{N}{F_s} B^{-1}\left(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1}\right) \tag{2-19}$$

where f_l and f_h are the lowest and highest frequencies of the filterbank in Hz, F_s the sampling frequency in Hz and N is the size of $S[k]$. B is the function that approximates the mel frequency scale,

$$B(f) = 1125\ln(1 + \frac{f}{700}) \tag{2-20}$$

and B^{-1} stands for its inverse. In case of $f_l = 0$, $f_h = 8000$ and $M = 24$, the resulting Mel filterbank is shown in Figure (**2-8**). Similar to previous section, the time–frequency information is represented by the scalar valued logarithmic energy features $x(t + d, f_k) \in \mathbb{R}$, where the set $\{f_k : k = 1, \ldots, N_f\}$ appraises the $N_f = 24$ frequency components. A resulting matrix of acoustic log-energy features $\boldsymbol{X}_t \in \mathbb{R}^{N_t \times N_f}$ is obtained for each window analysis at the time position t. Same notation $\boldsymbol{X}_t \in \mathbb{R}^{N_t \times N_f}$, in respect to mel-like WP feature estimation process, is used.

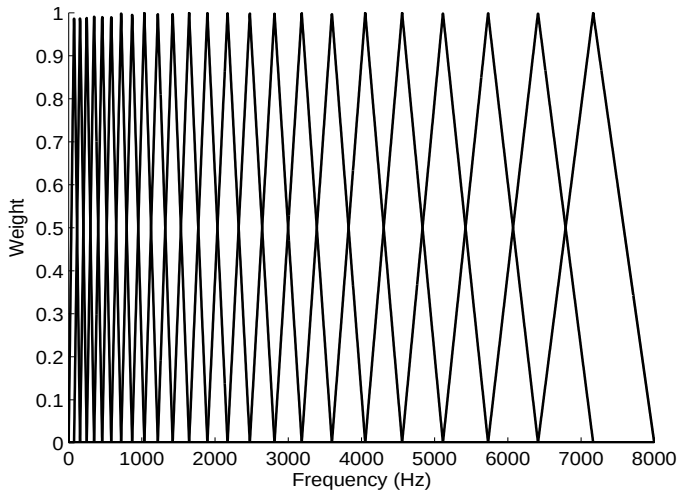


Figure 2-8: Distribution of triangular Mel-filters in the frequency range from zero upto 8000 Hz.

filter k	lower cut-off (Hz)	higher cut-off (Hz)	node W_l^p
1	0	125	W_6^0
2	125	250	W_6^1
3	250	375	W_6^2
4	375	500	W_6^3
5	500	625	W_6^4
6	625	750	W_6^5
7	750	875	W_6^6
8	875	1000	W_6^7
9	1000	1125	W_6^8
10	1125	1250	W_6^9
11	1250	1375	W_6^{10}
12	1375	1500	W_6^{11}
13	1500	1750	W_5^6
14	1750	2000	W_5^7
15	2000	2250	W_5^8
16	2250	2500	W_5^9
17	2500	2750	W_5^{10}
18	2750	3000	W_5^{11}
19	3000	3500	W_4^6
20	3500	4000	W_4^7
21	4000	5000	W_3^4
22	5000	6000	W_3^5
23	6000	7000	W_3^6
24	7000	8000	W_3^7

Table 2-2: Wavelet packet nodes associated to the mel–like filter banks that are used in the present work.

3 Proposed methods for articulatory inversion

3.1. TF relevant features for inversion

3.1.1. Measures of statistical association

Kendall coefficient

Given a bivariate distribution model of $x(t+d, f_k)$ and $y^m(t)$ random variables, the Kendall coefficient, noted τ, is also used as a measure of random association, which is defined in terms of probability P as follows:

$$\begin{aligned}\tau =& P((x_i(t+d, f_k) - y_i^m(t))(x_j(t+d, f_k) - y_j^m(t)) > 0) \\ & - P((x_i(t+d, f_k) - y_i^m(t))(x_j(t+d, f_k) - y_j^m(t)) < 0)\end{aligned} \tag{3-1}$$

Both terms of $\tau \in [-1, 1]$, in (3-1) are estimated from the given set of independent observations pairs $(x_i(t+d, f_k), y_i^m(t))$, $(x_j(t+d, f_k), y_j^m(t))$, selected among N samples. So, the measure τ becomes 1 if there is a perfect concordance, i.e., if the direct relationship holds, $x_i(t+d, f_k) \lesseqgtr x_j(t+d, f_k)$ whenever $y_i^m(t) \lesseqgtr y_j^m(t)$. On the contrary, the measure of perfect discordance yields -1 meaning that the inverse relationship holds: $x_i(t+d, f_k) \lesseqgtr x_j(t+d, f_k)$ whenever $y_i^m(t) \gtreqless y_j^m(t)$. If neither concordant criterion nor discordant criterion is true, the measure between pairs will lie within the interval $(-1, 1)$.

Given the specific set of pairs $(x_i(t+d, f_k), y_i^m(t))$, $(x_j(t+d, f_k), y_j^m(t))$, the respective

elemental pair indicator of association measure $K_{ij} \in [-1, 1]$ is defined in Eq. (3-2) as:

$$K_{ij} = \text{sgn}(x_i(t+d, f_k) - y_i^m(t))\,(\,x_j(t+d, f_k) - y_j^n(t)) \tag{3-2}$$

where $\text{sgn}(\cdot)$ stands for the signum function. Then, the value of $\tau_{d,k}^m = \boldsymbol{E}\{K_{ij}\}$ denoting the Kendall coefficient at the time shift d, given the filter bank number k and the EMA channel m, is provided by following expected value:

$$\tau_{d,k}^m = \sum_{1 \le i < j \le N} \sum \frac{K_{ij}}{\binom{N}{2}} \tag{3-3}$$

The Kendall association measure between the articulatory and the acoustic data roughly shows how articulatory information is coded in the time and frequency components of the speech signal. However, the vocal tract shape inference is not commonly carried out using a single feature. An additional question to be addressed is how the articulatory information is distributed if using more than one input. To clarify this issue the partial rank correlation could be used. So, given a trivariate population where the marginal distributions of each variable are continuous, it is necessary to determine a measure of the association between $x(t+d, f_k)$ and $y^m(t)$ when term $x' = x(t+d', f^m)$, remains constant, with $f^m \neq f_k$, and $d \neq d'$. That is, there is the need for computing the additional information provided by a new feature $x(t+d, f_k)$ for the inference of $y^m(t)$, given the feature x'.

Based on the estimated τ values between those pairs of variables involved in the partial correlation calculation, the partial rank correlation coefficient $T_{x,y\boldsymbol{\cdot}x'}$, in case of Kendall measure, can be calculated as follows [18]:

$$T_{x,y\boldsymbol{\cdot}x'} = \frac{\tau_{xy} - \tau_{xx'}\tau_{yx'}}{((1-\tau_{xx'}^2)(1-\tau_{yx'}^2))^{1/2}} \tag{3-4}$$

χ^2 Information measure

Denoted measure $I(x(\cdot), y(\cdot)) \in \mathbb{R}$ holds the information content, with regard to the articulatory trajectory $y^m(t) \in \boldsymbol{y}_t$, of each individual acoustic feature $x(t+d, f_k)$, which describes the TF–atom at time $t+d$ and frequency f_k in the TF plane $\boldsymbol{X}_t$. Generally speaking, mutual information and the χ^2 information are measures regarded as the distance

between a joint probability distribution $P_{xy}(\cdot,\cdot)$ and the product of marginal distributions, $P_x(\cdot)$ and $P_y(\cdot)$. Instead of former measure which is widely used, this study prefers the latter for validating the Kendall coefficient as the χ^2 information measure can be implemented without carrying out an explicit estimation of the joint probability density function [108]. Estimation of the information content by means of the χ^2 measure is written as follows [72]:

$$I(x(t+d,f_k),y^m(t)) = \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{(P_{xy}(x(t+d,f_k),y^m(t)) - P_x(x(t+d,f_k))P_y(y^m(t)))^2}{P_x(x(t+d,f_k))P_y(y^m(t))} dxdy \tag{3-5}$$

The information content of (3-5) can be estimated based on the density ratio, denoted as $r^m_{d,k} = r(x(t+d,f_k),y^m(t))$, between the random variables $x(t+d,f_k)$ and $y^m(t)$, as suggested in [108]:

$$I^m_{d,k} = I(x(t+d,f_k),y^m(t)) = \int_{\mathbb{R}}\int_{\mathbb{R}} (r^m_{d,k}-1)^2 P_x(x(t+d,f_k))P_y(y^m(t))\,dx\,dy \tag{3-6}$$

where the term $r^m_{d,k} \in \mathbb{R}$ determined as follows:

$$r^m_{d,k} = \frac{P_{xy}(x(t+d,f_k),y^m(t))}{P_x(x(t+d,f_k))P_y(y^m(t))} \tag{3-7}$$

where density ratio function of (3-7) is estimated based on the linear model:

$$\widetilde{r}^m_{d,k} = \boldsymbol{\alpha}^\top \boldsymbol{\Phi}(x(t+d,f_k),y^m(t)),\ \boldsymbol{\alpha} = [\alpha_1\,\alpha_2\,,\ldots,\,\alpha_{n_b}]^\top \in \mathbb{R}^{n_b\times 1} \tag{3-8}$$

being $\boldsymbol{\Phi}$ the matrix comprising the n_b base functions:

$$\boldsymbol{\Phi}(x(t+d,f_k),y^m(t)) = [\phi_1(\cdot,\cdot),\,\phi_2(\cdot,\cdot),\,\cdots,\,\phi_{n_b}(\cdot,\cdot)]^\top \tag{3-9}$$

where notation $^\top$ stands for the transpose of a vector. The weighting parameters of vector $\boldsymbol{\alpha}$ are to be learned from data samples, while the matrix $\boldsymbol{\Phi}$ is based on Gaussian kernels, as developed in [49]. Additional details about the estimation of the density ratio function of (3-7) can be found in [108, 72].

3.1.2. Regression by using Gaussian mixture models

With nonlinear regression the experimenter may not have any idea as to the true model but simply belives that the model is nonlinear [91]. In order to obtain the model that maps the acoustic space to the articulatory space, techniques such as artificial neural networks and gaussian mixture models have been used.

The task at hand consists on searching the estimation $\widetilde{y}_t$ of the articulatory configuration y_t from the acoustic vector $\boldsymbol{v}_t \in \mathbb{R}^{p\times 1}$, comprising p selected acoustic input features at the time moment t, i.e, $\widetilde{y}_t^m = \boldsymbol{E}\{y^m|\boldsymbol{v}=\boldsymbol{v}_t\} = \int P(y_t^m|v=v_t)y_t^m dy_t$. We assume that $y^m, \boldsymbol{v}$ are jointly distributed with a probability density function $P(\boldsymbol{z}_t;\cdot)$; where, $\boldsymbol{z}_t$ is the joint vector $\boldsymbol{z}_t = [(\boldsymbol{v}_t)^\intercal, (\boldsymbol{y}_t^m)^\intercal]$ and $^\intercal$ denotes the transpose of the vector. Thus, $\boldsymbol{z_t} \in \mathbb{R}^{p+1}$.

Articulatory trajectories contain data whose statistical modelling requieres more than one simple gaussian function. In this case, more complex models are necessary; for example, finite mixture models of the form:

$$P(\boldsymbol{z}_t\cdot) = \sum_j^J \pi^j P(\boldsymbol{z}_t;\boldsymbol{\theta}^j) \tag{3-10}$$

$\boldsymbol{\theta}^j$ is the parameter vector describing $P(\boldsymbol{z}_t;\boldsymbol{\theta}^j)$. The mixture weights, π^j, are normalized positive scalars ($\sum_{j=1}^J \pi^j = 1$ and $\pi^j > 0$). This ensures that the mixture is a true probability density function [119]. In the gaussian mixture model of Eq. 3-10, $P(\boldsymbol{z}_t;\boldsymbol{\theta}^j)$ is the multivariate gaussian distribution shown in equation 3-11, for which $\boldsymbol{\theta}^j = \{\boldsymbol{\mu}_z^j, \boldsymbol{\Sigma}_z^j\}$,

$$\begin{aligned} P(\boldsymbol{z}_t;\boldsymbol{\theta}^j) &= \frac{1}{(2\pi)^{d/2}} \mid \boldsymbol{\Sigma}^j \mid^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{z}_t-\boldsymbol{\mu}^j)^\intercal(\boldsymbol{\Sigma}^j)^{-1}(\boldsymbol{z}_t-\boldsymbol{\mu}^j)\right) \\ &= \mathcal{N}(\boldsymbol{z}_t;\boldsymbol{\mu}_z^j,\boldsymbol{\Sigma}_z^j) \end{aligned} \tag{3-11}$$

By using a sufficient number of gaussians, and by adjusting their means and covariances as well as the coefficients in the linear combination, almost any continuous density can be approximated [11]. The complete Gaussian mixture density is parameterized by the weights, the mean vectors and the covariance matrices from all component densities which is represented by the notation,

$$\boldsymbol{\Theta} = \left\{\pi^j, \boldsymbol{\mu}_z^j, \boldsymbol{\Sigma}_z^j\right\}, \quad j = 1,\ldots,J \tag{3-12}$$

Once established the model structure, what remains is to determine the parameters of the same model. For the sake of estimating the parameter's model Θ, the optimization of any criterion is applied; among which, the *likelihood criterion* is one of the most commnonly used. Given a set of N observations, the *likelihood function* is defined as [114],

$$P(\boldsymbol{Z};\boldsymbol{\Theta}) = \prod_{t=1}^{N} P(\boldsymbol{z}_t;\boldsymbol{\Theta}) \tag{3-13}$$

which corresponds to the the probability of the observed data $\boldsymbol{Z} = \{\boldsymbol{z}_t; t = 1,\ldots,N\}$ under the model $P(\boldsymbol{z}_t;\boldsymbol{\Theta})$. When maximun likelihood estimates are of interest, it is usually convenient to use the logarithm of the likelihhod function, named the *log-likelihood function*, rather than the likelihood function itself [119]. Thus, the log-likelihood function is given by,

$$L(\boldsymbol{\Theta},\boldsymbol{Z}) = -\sum_{t=1}^{N} \log P(\boldsymbol{z}_t;\boldsymbol{\Theta}) \tag{3-14}$$

The task at hand consists of finding $\boldsymbol{\Theta}^*$ such that $L(\boldsymbol{\Theta},\boldsymbol{Z})$ in 3-14 is maximized, that is,

$$\boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Theta}} L(\boldsymbol{\Theta};\boldsymbol{Z}) \tag{3-15}$$

Because it is not possible to solve $\partial L/\partial\boldsymbol{\Theta} = 0$ explicitly for the parameters of the model, iterative schemes must be employed [115]. One approach for maximising the likelihood L is to use the procedure known as EM (expectation-maximisation) algorithm. The EM algorithm iteratively increases the likelihood of the model parameters by successive maximizations of an intermediate quantity. The particular EM procedure used to infere articulatory configuration is shown in algorithm 1.

The mean vector $\boldsymbol{\mu}_z^j$ and covariance matrix $\boldsymbol{\Sigma}_z^j$ of the j_{ht} mixture component are decomposed as follows,

$$\boldsymbol{\mu}_z^j = \begin{bmatrix} \boldsymbol{\mu}_v^j \\ \boldsymbol{\mu}_y^j \end{bmatrix} \qquad \boldsymbol{\Sigma}_z^j = \begin{bmatrix} \boldsymbol{\Sigma}_{vv}^j & \boldsymbol{\Sigma}_{vy}^j \\ \boldsymbol{\Sigma}_{yv}^j & \boldsymbol{\Sigma}_{yy}^j \end{bmatrix} \tag{3-22}$$

Once joint probability $P(\boldsymbol{z_t};\cdot)$ has been estimated, what remeians is to determine the conditional probability $P(\boldsymbol{y}|\boldsymbol{v};\cdot)$. As shown in [110, 77], the conditional probability can also

[1] Initialize $\boldsymbol{\mu}^j$, $\boldsymbol{\Sigma}^j$ and π^j. Evaluate the log likelihood.
while Stopping criterion is not reached **do**
[2] Evaluate the responsabilities using the current parameter values,

$$\beta^j(\boldsymbol{z}_t) = \frac{\pi^j \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_z^j, \boldsymbol{\Sigma}_z^j)}{\sum_{i=1}^{J} \pi^i \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_z^i, \boldsymbol{\Sigma}_z^i)} \tag{3-16}$$

[3] Estimate the parameters but using the current responsabilities,

$$\boldsymbol{\mu}_z^j = \frac{1}{N^j} \sum_{t=1}^{N} \beta^j(\boldsymbol{z}_t) \boldsymbol{z}_t \tag{3-17}$$

$$\boldsymbol{\Sigma}_z^j = \frac{1}{N^j} \sum_{t=1}^{N} \beta^j(\boldsymbol{z}_t)(\boldsymbol{z}_t - \boldsymbol{\mu}^j)(\boldsymbol{z}_t - \boldsymbol{\mu}^j)^\intercal \tag{3-18}$$

$$\pi^j = \frac{N^j}{N} \tag{3-19}$$

where

$$N_j = \sum_{t=1}^{N} \beta^j(\boldsymbol{z}_t) \tag{3-20}$$

[4] Evaluate the likelihood

$$\log P(\boldsymbol{Z}; \Theta) = \sum_{t=1}^{N} \log \left(\sum_{j=1}^{J} \pi^j \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}^j, \boldsymbol{\Sigma}^j) \right) \tag{3-21}$$

end while

Algorithm 1: Expectation-Maximization algorithm for parameters estimation of probability density function modelled by gaussian mixtures models.

be expressed as a GMM, as follows:

$$P(\boldsymbol{y}|\boldsymbol{v};\boldsymbol{\mu}^j_{y|v},\boldsymbol{\Sigma}^j_{y|v}) = \sum_{j=1}^{J}\beta^j(\boldsymbol{v}_t)\mathcal{N}(\boldsymbol{y};\boldsymbol{\mu}^{j,t}_{y|v},\boldsymbol{\Sigma}^j_{y|v}) \tag{3-23}$$

where,

$$\boldsymbol{\mu}^{j,t}_{y|v} = \boldsymbol{\mu}^j_y + \boldsymbol{\Sigma}^j_{yv}(\boldsymbol{\Sigma}^j_v)^{-1}(\boldsymbol{v}_t - \boldsymbol{\mu}^j_v) \tag{3-24}$$

is the conditional mean; and,

$$\boldsymbol{\Sigma}^j_{y|v} = \boldsymbol{\Sigma}^j_{yy} - \boldsymbol{\Sigma}^j_{yv}(\boldsymbol{\Sigma}^j_{vv})^{-1}\boldsymbol{\Sigma}^j_{yv} \tag{3-25}$$

is the conditional covariance. $\beta^j(\boldsymbol{v}_t)$ is computed by using the following expression:

$$\beta^j(\boldsymbol{v}_t) = \frac{\pi^j\mathcal{N}(\boldsymbol{v}_t;\boldsymbol{\mu}^j_v,\boldsymbol{\Sigma}^j_v)}{\sum_{i=1}^{J}\pi^i\mathcal{N}(\boldsymbol{v}_t;\boldsymbol{\mu}^i_v,\boldsymbol{\Sigma}^i_v)} \tag{3-26}$$

Lastly, the estimation $\widetilde{\boldsymbol{y}}_t$ yields

$$\widetilde{\boldsymbol{y}}_t = \sum_{j=1}^{J}\beta^j(\boldsymbol{v}_t)(\boldsymbol{\mu}^j_v + \boldsymbol{\Sigma}^j_{yv}(\boldsymbol{\Sigma}^j_{vv})^{-1}(\boldsymbol{v}_t - \boldsymbol{\mu}^j_v)) \tag{3-27}$$

3.2. Vocal tract modeling

Vocal tract models, also called articulatory synthesizers, generate speech sounds from vectors contaning information about articulatory configuration. An articulatory synthesizer is divided into two parts, see figure **3-1**. First, articulatory models are usually used to map articulatory parameters to an estimated geometric shape of the vocal tract from which the cross-sectional area function of the vocal tract can be specified [94, 12].

One of the most widely used articulatory model is Maeda's model [76], which is presented in [67, 69, 70]. Second, the acoustic model is a non-uniform acoustic tube that produces the acoustic outcomes by using the cross-sectional areas provided by the articulatory model. A time-domain simulation method of the vocal-tract system is described in [68].

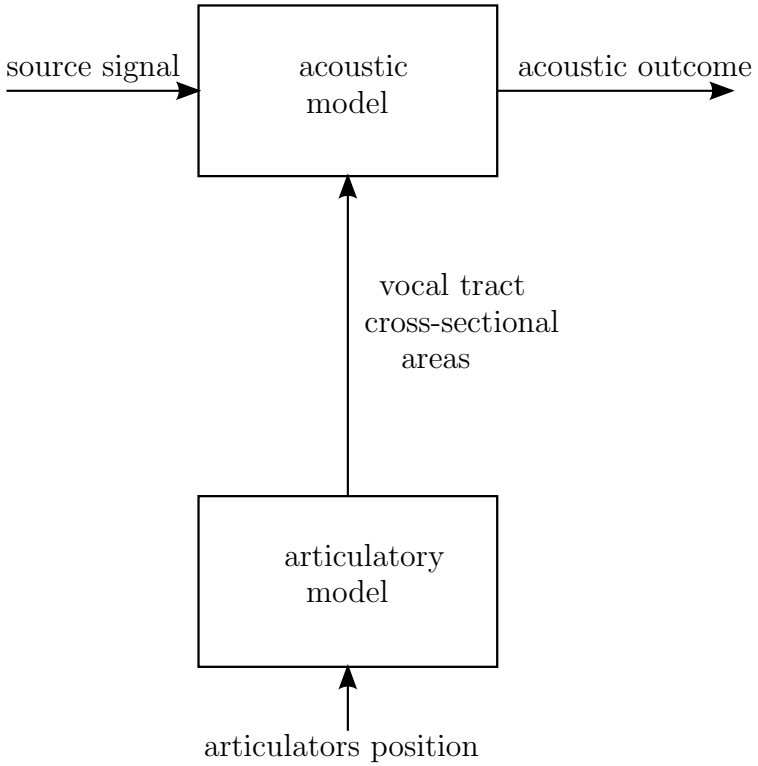


Figure 3-1: Schematic model of an articulatory synthesizer.

3.2.1. Acoustic modeling

The model describing the vocal tract behavior is complicated and impractical. To simplify the model, we assume that sound waves obey planar propagation along the axis of the vocal tract. This assumption is valid only for frequencies less than approximately 4 KHz. In such case, only the cross-sectional area and the perimeter along the length of the vocal tract determine its acoustic characteristics [68]. The equations that characterize the volume velocity $\mathrm{u}(x,t)$ and sound pressure $p(x,t)$ along the vocal tract from glottis to lips can be described as follows [16],

$$-\frac{\partial \rho(\varkappa,t)}{\partial \varkappa} = \sigma \frac{\partial(\frac{\mathrm{u}(\varkappa,t)}{A(\varkappa,t)})}{\partial t} \tag{3-28}$$

$$-\frac{\partial \mathrm{u}(\varkappa,t)}{\partial \varkappa} = \frac{1}{\sigma c^2}\frac{\partial(\frac{\rho(\varkappa,t)}{A(\varkappa,t)})}{\partial t} + \frac{\partial A(\varkappa,t)}{\partial t} \tag{3-29}$$

where,

$$\begin{aligned}
\rho(\varkappa,t) &\quad \text{sound pressure}\\
\mathrm{u}(\varkappa,t) &\quad \text{volume velocity of air}\\
A(\varkappa,t) &\quad \text{cross-sectional area}\\
\sigma &\quad \text{density of air inside the vocal tract}\\
c &\quad \text{speed of sound in air}\\
\varkappa &\quad \text{distance from glottis}\\
t &\quad \text{time}
\end{aligned} \tag{3-30}$$

The model just shown can be approximated by a set of linear differential equations. The pressure $\rho(\varkappa,t)$ and the volume velocity $u(\varkappa,t)$ inside an acoustic tube with non-rigid walls are governed, in the first order approximation, by the following three partial differential equations [68]: the equation of motion (EMo), that of continuity (EC), and wall vibration (EV), as

$$\frac{\partial\rho(\varkappa,t)}{\partial\varkappa}+\frac{\partial}{\partial}\frac{\sigma_0 u(\varkappa,t)}{A_0}+\frac{r\mathrm{u}(\varkappa,t)}{A_0(\varkappa,t)}=0 \tag{3-31}$$

$$\frac{\mathrm{u}(\varkappa,t)}{\partial\varkappa}+\frac{\partial}{\partial t}\frac{A_0\rho(\varkappa,t)}{\sigma_0 c^2}+\frac{\partial A_0(\varkappa,t)}{\partial t}+\frac{\partial S_0 y}{\partial t}=0 \tag{3-32}$$

$$M_a\frac{\partial^2 y}{\partial t^2}+B_a\frac{\partial y}{\partial t}+K_a y=S_0\rho(\varkappa,t) \tag{3-33}$$

These equations represent the acoustic transmission line of the vocal tract. A given area function is denoted by $A_0=A_0(\varkappa,t)$, which is related to the previously defined area function $A(\varkappa,t)$ by

$$A(\varkappa,t)=A_0(\varkappa,t)+y(\varkappa,t)S_0(\varkappa,t) \tag{3-34}$$

where $S_0(\varkappa,t)$ indicates a given parameter of the tube, and $y(\varkappa,t)$ the amplitude of the yielding of walls due to the sound pressure inside the tube. The coefficients, M_a, B_a and K_a in Eq. 3-33 represents respectively the mass, mechanical resistance, and the stiffness of the wall per unit length of the tube. These coefficients are assumed to be constant and uniform

along the vocal tract, even though their actual values vary depending on the location and also on the tenseness of the muscles beneath the wall surface [68].

As shown in the model depicted in figure **3-3**, the glottal end of the pharyngeal tube is directly connected to the pressure source. The boundary condition at that end is given by,

$$P_{sub}(t) = \rho(\varkappa_0, t) \tag{3-35}$$

where P_{sub} indicates a given subglottal air pressure, and $\varkappa_0$ is the coordinate value at that end.

At the nasal coupling point, whose location is defined as $\varkappa = \varkappa_k$ in figue **3-3**, the volume velocity and the pressure must satisfy the following conditions,

$$\mathrm{u}(\varkappa_k^-, t) = \mathrm{u}(\varkappa_k^+, t) + \mathrm{u}'(t, 0) \tag{3-36}$$

$$\rho(\varkappa_k^-, t) = p(\varkappa_k^+, t) = \rho'(t, 0) \tag{3-37}$$

where the superscript $-$ indicates the pharyngeal end, and $+$ stands for the inlet of the oral cavity. $'$ refers to the nasal tract.

In order to obtain a numerical simulation of the continous equations $\rho(x,t)$ y $\mathrm{u}(\varkappa,t)$ a discretization process in time and space is performed [68]. X_j denoted discrete values of $\varkappa$. The discrete value of ρ (P_j) is the central pressure of the j_{th} section delimited by the points X_{j-1} and X_j. Once these approximations are made, the resulting equation of motion (EMo), that of continuity (EC), and the equation of wall vibration (EV) take the form,

$$P_{j-1} - P_j = \frac{d}{dt}\frac{\sigma_0 X_{j-1}}{2A_{j-1}}\mathrm{U}_j + \frac{X_{j-1} r_{j-1}}{2A_{j-1}}\mathrm{U}_j + \frac{d}{dt}\frac{\sigma_0 X_j}{2A_j}\mathrm{U}_j + \frac{X_j r_j}{2A_j}\mathrm{U}_j \tag{3-38}$$

$$\mathrm{U}_j - \mathrm{U}_{j+1} = \frac{d}{dt}\frac{X_j A_j}{\sigma_0 c^2}P_j^2 + \frac{d}{dt}X_j A_j + \frac{d}{dt}X_j S_j y_j \tag{3-39}$$

$$S_j P_j = m\frac{d^2}{dt^2}y_j + b\frac{d}{dt}y_j + k y_j \tag{3-40}$$

The equations 3-39 and 3-40 are combined by eliminating y_j, as folows;

$$\mathrm{U}_j - \mathrm{U}_{j+1} = \mathrm{u}_1 + \mathrm{u}_2 + \mathrm{u}_3 \tag{3-41}$$

$$P_j = \frac{d}{dt}\frac{m}{X_j S_j^2}\mathrm{u}_1 + \frac{b}{X_j S_j^2}\mathrm{u}_2 + \int_0^t \frac{k}{X_j S_j^2}\mathrm{u}_3 dt \tag{3-42}$$

where,

$$\begin{aligned} \mathrm{u}_1 &= \frac{d}{dt}\frac{X_j A_j}{\sigma_0 c^2} P_j \\ \mathrm{u}_2 &= \frac{d}{dt} X_j A_j \\ \mathrm{u}_3 &= \frac{d}{dt} X_j S_j y_j \end{aligned} \tag{3-43}$$

Now let us show that the above equations lead to the lumped transmission-line approximation of the vocal tract in which the tract is represented by a series of symmetric T-type electrical networks. Where, for each cross-sectional area, the variables P and U corresponds to voltage and electrical current, respectively. The electrical analogy is shown in figure 3.2.1. The electrical components are defined as

$$\begin{aligned} L_j &= \frac{\sigma_0 X_j}{A_j} \\ R_j &= \frac{4\pi\mu X_j}{A_j} \\ C_j d_j &= -\frac{d}{dt}(X_j A_j) \\ Lw_j &= \frac{m}{X_j S_j^2} \\ Cw_j &= \frac{X_j S_j^2}{k} \end{aligned} \tag{3-44}$$

Using the notation defined in Eq. 3-44, the set of equations are rewritten in more compact form as

$$P_{j-1} - P_j = \frac{d}{dt}(L_{j-1})\mathrm{U}_j + (R_{j-1} + R_j)\mathrm{U}_j \tag{3-45}$$

$$\mathrm{U}_j - \mathrm{U}_{j-1} = \mathrm{u}_3 + \mathrm{u}_3 + \mathrm{u}_3 \tag{3-46}$$

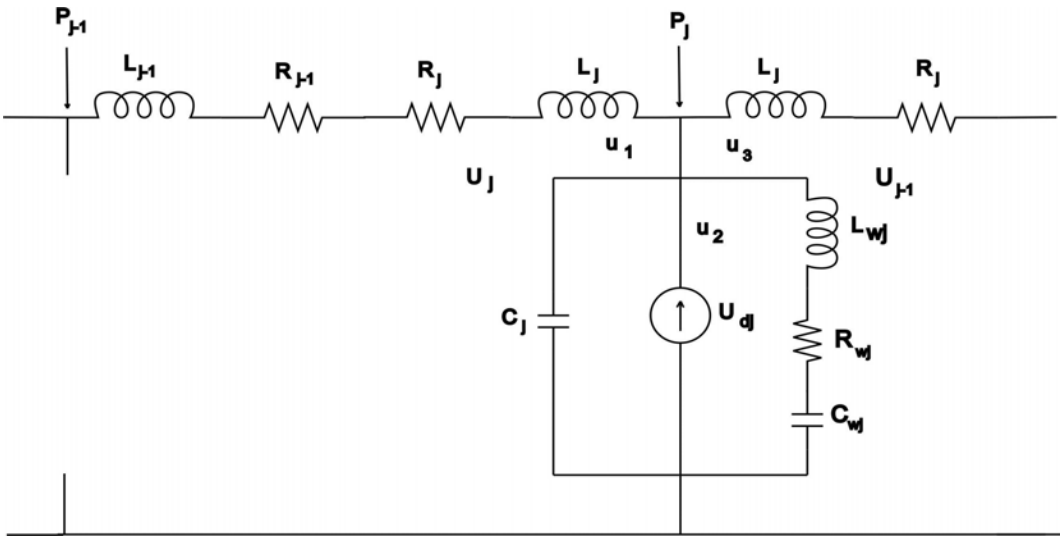


Figure 3-2: Electrical analogy of a cross-sectional area.

where,

$$\mathrm{u}_1 = \frac{d}{dt} C_j P_j \tag{3-47}$$

$$\mathrm{u}_2 = -\mathrm{U}d_j \tag{3-48}$$

$$P_j = \frac{d}{dt} Lw_j \mathrm{u}_3 + Rw_j \mathrm{u}_3 + \int_0^t \frac{\mathrm{u}_3}{Cw_j} dt \tag{3-49}$$

The resulting lumped transmission-line approximation is depicted in figure 3.2.1. L_g, R_g and P_{Sub} form the source that feeds the vocal tract. S_R - G_R and $S_R^{'}$ - $G_R^{'}$ represents the radiation at lips and nostrils, respectively. P_K in figure **3-3** is the electrical approximation for each of the cross-sectional areas.

3.2.2. Articulatory model

In the articualtory model proposed in [67, 69, 70], the vocal tract shapes are assumed to be functions of features (elementary articulators) such as jaw position, vertical movement; tongue dorsum position that can move roughly horizontally from the front to the back of the mouth cavity; tongue dorsum shape; apex position, this parameter only deforms the apex part of the tongue by moving it up or down; lip height; lip protrusion; and, larynx height

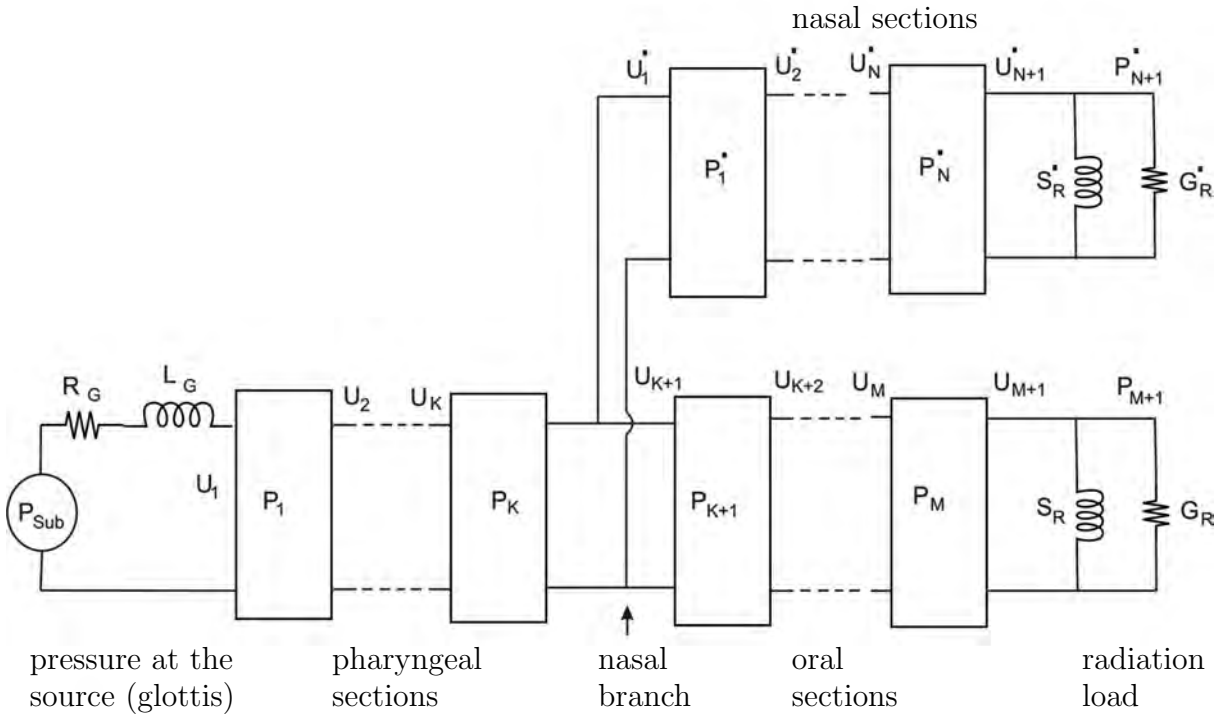


Figure 3-3: The articulatory synthesizer model.

[82]. Such parameters are depicted in figure **3-4**.

The vocal tract contour is represented by the points resulting from the intersection between the semipolar lines and the tract shape, see figure **3-5**. The distances between these intersection points form the representation vector of the internal vocal tract ζ_t. The dimension of the representation vector depends on the number of grid lines. Vector ζ_t can be represented by using linear components ξ as follows,

$$\zeta_t = \Lambda \, \xi_t \tag{3-50}$$

where ξ_t is the set of elementary articulators; and, Λ is a matrix of weigths saving the influence of the factors.

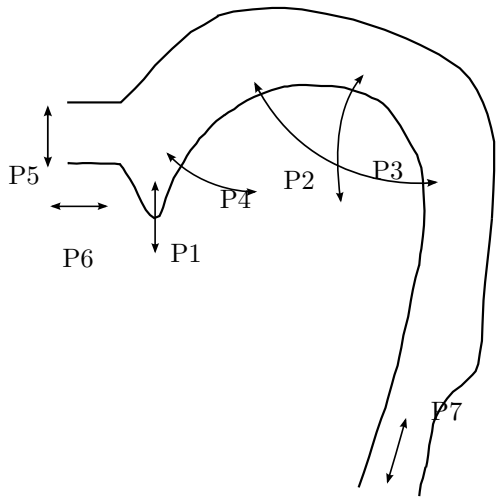


Figure 3-4: Maeda's parameters: P1, jaw position; P2, tongue dorsum position; P3, tongue dorsum shape; P4, apex position; P5, lip height; P6, lip protrusion; P7, larynx height. Credits: Blaise Potard, see [82].

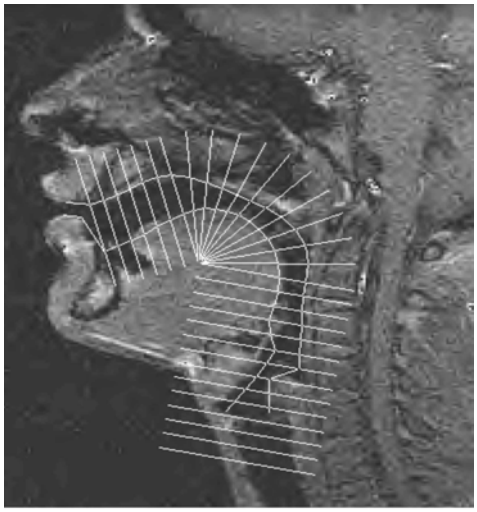
3.3. Weighted cepstral distance learning for accessing articulatory codebooks

3.3.1. Cepstral distance measure

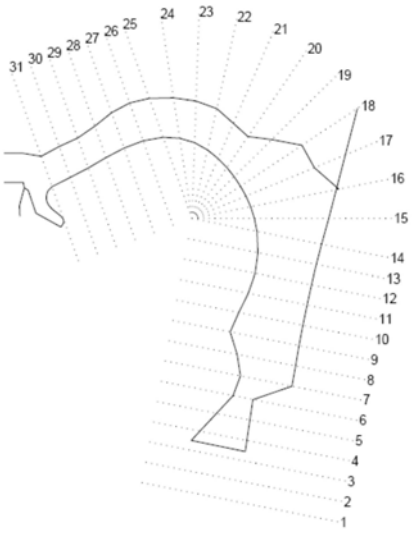
The cesptrum of a discrete-time signal is defined as the inverse Fourier transform of the logarithm of its power spectrum [16]. For speech processing a distance measure $\mathrm{d}(\mathbf{c}, \hat{\mathbf{c}})$, where $\mathbf{c}$ and $\hat{\mathbf{c}}$ are two occurences of a particular acoustic representation, should satisfy at least the following two properties [54, 33]:

$$\begin{aligned} &\text{symmetry} && \mathrm{d}(\mathbf{c}, \hat{\mathbf{c}}) = \mathrm{d}(\hat{\mathbf{c}}, \mathbf{c}) \\ &\text{positive definiteness} && \begin{cases} \mathrm{d}(\mathbf{c}, \hat{\mathbf{c}}) > 0 & \forall\ \mathbf{c} \neq \hat{\mathbf{c}} \\ \mathrm{d}(\mathbf{c}, \hat{\mathbf{c}}) = 0 & \text{if } \mathbf{c} = \hat{\mathbf{c}} \end{cases} \end{aligned} \tag{3-51}$$

The distance from $\mathbf{c}$ to $\hat{\mathbf{c}}$ is equal to the distance from $\hat{\mathbf{c}}$ to $\mathbf{c}$, and a distance between $\mathbf{c}$ and $\hat{\mathbf{c}}$ is positive, except for the case where $\mathbf{c} = \hat{\mathbf{c}}$; in this case the distance is zero [54]. The symmetry requirement insures that a distance measure between portions of sounds does not distinguish between which is a reference and which is a test sound [33]. $\mathbb{L}^p$ distances are



(a) MRI image and vocal tract contour



(b) Semipolar system used to parametrize the vocal tract contour

Figure 3-5: Vocal tract parametrization. Credits: Yves Laprie, see the document *Introduction à lácoustique de la parole* in `http://www.loria.fr/~laprie/`.

commonly used in engineering, and they are defined by the expression [47]:

$$\mathrm{d}(\mathbf{c}, \hat{\mathbf{c}}) = \left(\sum_k \mid c_k - \hat{c}_k \mid^p \right)^{\frac{1}{p}} \tag{3-52}$$

where c_k and $\hat{c}_k$ are the entries of the vectors c_k and $\hat{c}_k$, respectively. The $\mathbb{L}^p$ measures, for $p \geq 1$, are true metrics in the sense that they satisfy the triangle inequality $\mathrm{d}(\mathbf{c}, \hat{\mathbf{c}}) \leq \mathrm{d}(\mathbf{c}, \mathbf{c}') + \mathrm{d}(\mathbf{c}', \hat{\mathbf{c}})$ [54], in addition to being symmetric and positive definite.

The equation

$$\mathrm{d}^2(\mathbf{c}, \hat{\mathbf{c}}) = (\mathbf{c} - \hat{\mathbf{c}})^t \mathbb{W}^t \mathbb{W} (\mathbf{c} - \hat{\mathbf{c}}) \tag{3-53}$$

corresponds to the $\mathbb{L}^2$ or Euclidean distance between the linearly transformed versions $\mathbb{W}\mathbf{c}$ and $\mathbb{W}\hat{\mathbf{c}}$ of the elements $\mathbf{c}$ and $\hat{\mathbf{c}}$ respectively. The selection of a diagonal matrix for $\mathbb{W}$ gives rise to the distance,

$$\mathrm{d}^2(\mathbf{c}, \hat{\mathbf{c}}) = \sum_k w_k^2 (c_k - \hat{c}_k)^2 \tag{3-54}$$

where w_k are the components in the principal diagonal of $\mathbb{W}$.

If we take a truncated series to define a cepstral measure d_L as

$$\mathrm{d}_L^2(\mathbf{c}, \hat{\mathbf{c}}) = \sum_{k=1}^{L} w_k^2 (c_k - \hat{c}_k)^2 \ , \tag{3-55}$$

the distance d_L can be interpreted as the RMS (root mean square) distance between the log-spectra after each log-spectrum has been cepstrally smoothed to L coefficients [95]. The process described by Eq. (3-55) is equivalent to the liftering process used in several applications, and in this work we use this approach to alleviate problems caused by glottal variability. Constant, Juang and Meyer lifters have been used in speech-related tasks [45, 74, 111]. The constant lifter is given by

$$w_k = 1, \quad \text{for} \quad 0 < k \leq L; \tag{3-56}$$

the Juang lifter by

$$w_k = 1 + \frac{1}{2} k_m \sin \frac{k\pi}{L} \quad \text{for} \quad 0 < k \leq L; \tag{3-57}$$

and the Meyer lifter by

$$w_k = \begin{cases} (k/20)^{0,4} & \text{for} \quad 0 < k \leq 20 \\ 0{,}5 + 0{,}5 \cos(\frac{\pi(k-21)}{20}) & \text{for} \quad 20 < k \leq L. \end{cases} \tag{3-58}$$

The shapes of the Juang and Meyer lifters are depicted in Fig. **3-6**.

3.3.2. Cost function

Formant frequencies can change within distinct, frequency-dependent, boundaries with no perceptual effects. These boundaries are specified by the concept of Just Noticable Differences (JND). Based on the work of Ghitza and Goldstein [29], Schroeter et al. [93] provide the following expression:

$$JND\ \mathrm{F}_i(\,\%) = \begin{cases} 14{,}1\mathrm{F}_i^{-0,41} & 0 < \mathrm{F}_i \leq 1{,}5 \text{ kHz} \\ 6{,}3\mathrm{F}_i^{1,58} & 1{,}5 < \mathrm{F}_i \leq 4{,}0 \text{ kHz} \end{cases} \tag{3-59}$$

where F_i is the i_{th} formant value in kHz.

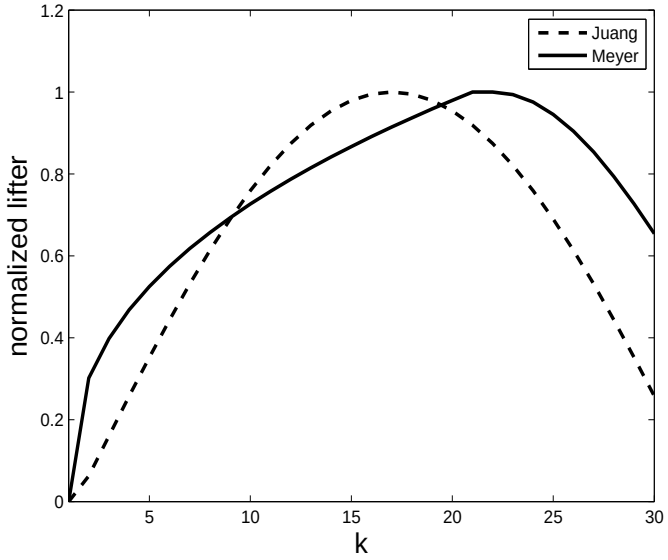


Figure 3-6: Shapes of Juang and Meyer lifters.

We calculate the weighted cepstral distance, as given by Eq. (3-55), of every frame in the speech signal to all the entries in the codebook. The codebook entries with the minimal cepstral distances are then retrieved, and the corresponding formants are compared with the measured formants of the real speech frame by means of the JND expression of Eq. (3-59). The cost function we use to optimize the weighting matrix $\mathbb{W}$ has two versions.

The cost function is similar to that used in [93]. In this case we retrieve from the codebook the ten entries with the smallest weighted cepstral distances, and of these we select the one with the minimal value from the perspective of the JND measure. That is, we demand that at least one entry among ten, is inside the JND boundaries. The process is repeated for all speech frames. The cost function is defined as the percentage of speech frames for which the codebook derived formants lie outside the range provided by the JND expression, as calculated for the corresponding real speech formants.

The cost function is defined as

$$O_i^{\mathrm{F}}(\,\%) = \frac{n_i^{\mathrm{F}}}{N_r} 100, \tag{3-60}$$

where N_r the number of frames used for the evaluation, and n_i^F is the number of frames with

$$100 \frac{|\,\mathrm{F}_i - \hat{\mathrm{F}}_i\,|}{\mathrm{F}_i} > JND(\mathrm{F}_i). \tag{3-61}$$

The final equation used to evaluate the w_k coefficients is obtained by averaging the last expression over the three formants frequencies:

$$O^{\bar{\mathrm{F}}}_{1-3} = \frac{1}{3}\sum_{i=1}^{3} O^{\mathrm{F}}_i. \tag{3-62}$$

3.3.3. Optimization of the cost function

Our goal is to find an optimal set of weights w_k, such that the codebook evaluation function $O^{\bar{\mathrm{F}}}_{1-3}$ of Eq. (3-62) is minimized. The look-up table procedure and the way that every selected entry is evaluated makes codebook access a nonlinear programming task. The appropriate optimization method to be used for such problems depends on the properties of the cost function (e.g. convexity, continuity and derivative related properties). Even though optimization methods based on derivative terms are faster, it is necessary that the first and/or second derivative terms exist. In our cost function, the first derivative term is not continuous because of the hard threshold caused by applying the inside-outside JND range procedure in Eq. (3-61). In addition, the JND expression of Eq. (3-59) does not have a continuous first derivative value over the whole range of formant evaluation.

Many stochastic techniques have been proposed in the literature to optimize this kind of cost functions, but, most of them display a practical behavior which inhibits their use for dimensions greater than, say, 10 or 15 [92]. It is known as the curse-of-dimensionality problem. For example, the performance of evolutionary algorithms deteriorates rapidly as the dimensionality of the search space increases [118]. In our case, i.e. searching for 30 lifter coefficients, it would be necessary to select a population of at least 60 individuals per generation (a considerable part of them randomly) and, given that evolutionary algorithms have a slow rate of convergence, it would be necessary to perform many evaluations. For the case of grid based methods, constructing a grid with ten steps per dimension needs about 10^{30} evaluations, what turns out to be almost impossible to solve from a practical point of view.

To find the optimal lifter we have to deal with two drawbacks: 1) the high dimensionality nature of the problem and, 2) the fact that the cost function is non-smooth, non-convex and has a considerable number of local minima. The lack of derivative information leads us to the use of some kind of *direct search method*, whose rate of convergence is slower [53]. Optimization in a high dimensional space without using derivative terms necessitates the use of a large number of cost function evaluations whch are themselves time-consuming.

The optimization is performed in two stages: first, a representation of the liftering function with a smaller number of parameters is found and, once the problem has been transformed to a problem of lower dimensionality, the Nelder-Mead optimization method is applied. However, this method is able to find local, rather than global, minima. To solve this problem the algorithm is initialized around zones containing relevant minima. Even though grid methods are not feasible for high dimensional problems they can be conveniently used for problems with small number of variables, to provide a good initialization point for a more efficient method [87]. This is the strategy we use in the present work.

Liftering function To avoid the curse-of-dimensionality problem, we use the cepstral smoothing function proposed for speech recognition by Itakura and Umezaki [42]. The authors of that work noted that the variance of the higher cepstral coefficients was much smaller than the variance of the lower cepstral coefficients. They found correlations between the k^2 terms and the inverse variance of the cepstral coefficients. This suggests that $w_k^2 = k^2$ (for $k = 1, 2, \ldots, L$) could be used to equalize the contributions for each term to the square cepstrum difference. The linear lifter $w_k = k$ suppresses the lower order quefrency components of the cepstrum in the time domain and enhances the higher order quefrency components. However, when using the linear lifter, the small variations in the spectrum are overemphasized due to the enhancement of the cepstral components at large enough values of k. Thus, to a reasonable extent, suppression of higher order cepstral components is necessary.

Rabiner and Schafer [86] describe the function

$$w_k = k^s \, e^{\frac{-k^2}{2T^2}} \quad k = 1, 2, \ldots, L. \tag{3-63}$$

where the term $e^{\frac{-k^2}{2T^2}}$ suppresses the cepstral components at large values of k.

Fig. **3-7** shows the configurations of the lifter components w_k, when T is varied from 5 to 30 for the values $s = 1, 2$ in Eq. (3-63). As s increases, lower order cepstral components are suppressed. As T increases, the weights for the higher order cepstral components increase. Even though this function cannot fully span the space $\mathbb{R}^L$ of all possible lifters, it may give rise to efficient lifters if values for s and T are optimally selected.

Nelder-Mead algorithm The Nelder-Mead algorithm [57], is a free-derivative search method, which only needs the evaluation of the objective function. The method uses a simplex program of $n+1$ points for n-dimensional vectors. In 2-dimensional space, a simplex iteration is

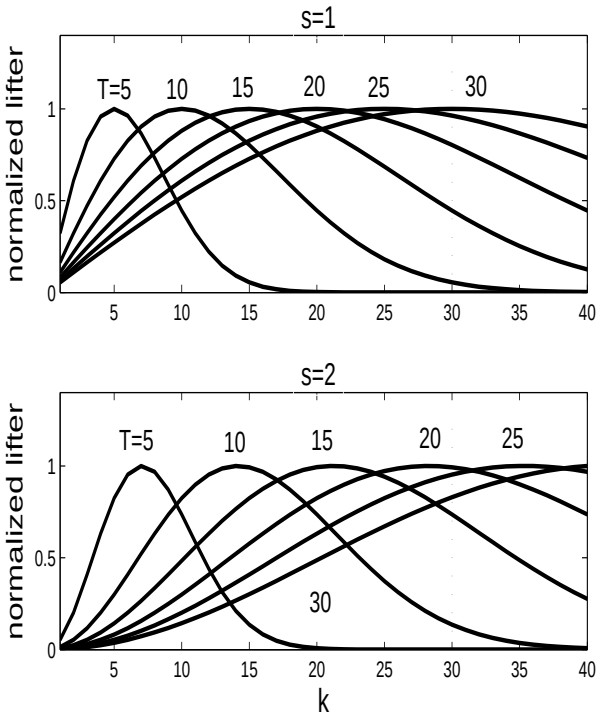


Figure 3-7: Lifter shapes generated from the function of Eq. 3-63 for various values for the parameters s and T.

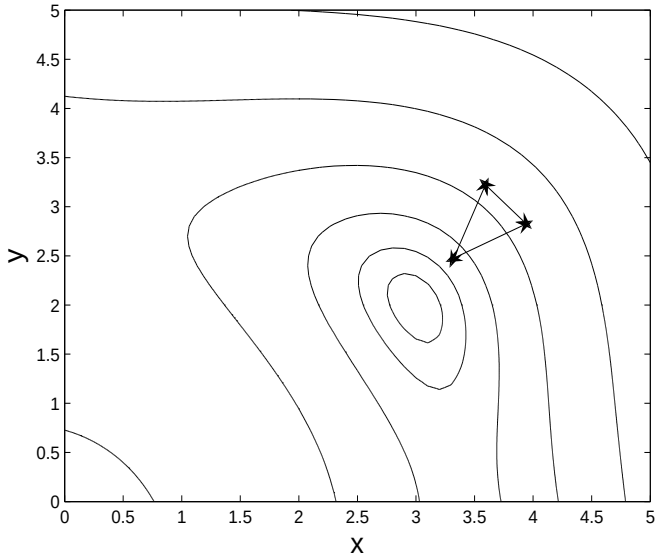


Figure 3-8: Schematic of the Nedler-Mead algorithm. For the 2-dimensional case, a simplex is defined as a triangle. Each simplex vertex corresponds to a possible solution. The contours in this example belong to the function $f(x, y) = \log((x^2 + y - 11)^2 + (x + y^2 - 7)^2 + 1)$.

a triangle determined by three points (vertices) and their interconnecting line segments. At every point the objective function is evaluated. The point with the highest numerical value of all three points is perpendicularly reflected against the opposite plain segment. This is called a reflection. The reflection can be accompanied with an expansion to take larger steps or with a contraction to shrink the simplex where an optimization valley floor is reached. The optimization procedure continues until some termination criteria are met. In this study we stop when a maximum number of iterations is reached. Fig. **3-8** depicts one iteration of the algorithm.

4 Experimental setup

4.1. Testing cepstral distance measures

Database

The database is formed by two components: real speech frames and synthesized speech data, with the latter stored in a codebook.

Codebook Articulatory codebooks are intended to obtain local approximations of the acoustic-to-articulatory mapping in a fast way. This is usually done by computing acoustic images of some articulatory vectors, and then interpolating the rest of the articulatory-to-acoustic mapping from these vectors. The codebook we use is a hypercube codebook. The characteristics of this codebook can be found in [81, 82]. The codebook itself is constructed using a recursive exploration of the articulatory space, partitioning it into hypercubes where the articulatory-to-acoustic mapping is pseudo-linear with a homogeneous acoustic error. Overall, this codebook contains a piece-wise linear approximation of the articulatory-to-acoustic mapping, each piece being a 7-hypercube, i.e. the generalization of a rectangle in a 7-dimensional space.

The codebook is used to organize the acoustic-articulatory information. The data used in that work comes from the speaker PB in [82] whose X-ray images were used to create the articulatory model by Maeda. The parameters of the articulatory model was achived by using the method developed in [44].

A hypercube H_c is characterized by the coordinates of its center P_0 and by the N-dimensional vector $\overrightarrow{r}$ of its length along each dimension, according to the following formula:

$$H_c(P_0, \overrightarrow{r}) = x \in \mathbb{R}^N \mid \forall i \in 1, ..., N, \mid (x - P_0)_i \mid \leq r_i \tag{4-1}$$

In such a region, an approximation of the acoustic image of an articulatory vector P_x belonging to H_c is computed using a linear approximation at the center of H_c, i. e. using the formula:

$$f^*(P_x) = F_0 + J_f(P_0)(P_x - P_0) \tag{4-2}$$

where F_0 is the acoustic image (a vector consisting of the first three formants) of the center P_0, and $J_f(P_0)$ is a Jacobian matrix of the articulatory to acoustic mapping computed around P_0. Therefore, each hypercube is characterized by its center P_0, its length vector $\overrightarrow{r}$, and its acoustic image by F_0 and $J_f(P_0)$.

Acoustic speech representation A series of recordings of vowel-vowel sequences, uttered by a male speaker, are used for the experiments in the present work. The speech signal was originally recorded at 44100 Hz, and then subsampled to 20000 Hz, in agreement with the sampling frequency of the synthesized signals in the codebook. Formant estimation is carried out using the automatic formant tracking algorithm described in [58], edited manually when needed. The dataset was composed by the estimations for 518 frames. 400 frames are used for training while 118 frames are used for testing. 30 cepstral coefficients were computed around the time instants in the speech signal where formants were estimated using windows of 512 points and FFT order of 512.

Microphones, room acoustics and distance from the microphone cause random disturbances in the speech recording [16]. In order to reduce these effects, the estimated cesptral parameters were transformed so that they have the same mean and standard deviation values as the synthesized ones. The real cepstrum $\mathbf{c}$ is transformed to $\mathbf{c}''$ by the process described in Eqs. (4-3) and (4-4).

$$\mathbf{c}' = \frac{\mathbf{c} - \mu_r}{\sigma_r} \tag{4-3}$$

$$\mathbf{c}'' = \sigma_s(\mathbf{c}' + \mu_s) \tag{4-4}$$

where μ_r, σ_r, μ_s and σ_s are defined by the expressions,

$$\mu_r = \frac{1}{N_r}\sum_{n}^{N_r}\overline{\mathbf{c}^n} \tag{4-5}$$

$$\sigma_r = \frac{1}{N_r}\sum_{n=1}^{N_r}\left(\sqrt{\frac{1}{L}\sum_{k=1}^{L}(c_k - \overline{\mathbf{c}^n})^2}\right) \tag{4-6}$$

$$\mu_s = \frac{1}{N_s}\sum_{n}^{N_s}\overline{\hat{\mathbf{c}}^n} \tag{4-7}$$

$$\sigma_r = \frac{1}{N_s}\sum_{n=1}^{N_s}\left(\sqrt{\frac{1}{L}\sum_{k=1}^{L}(c_k - \overline{\hat{\mathbf{c}}^n})^2}\right) \tag{4-8}$$

where $\mathbf{c}$ and $\hat{\mathbf{c}}$ are the cepstra of the real frame and the synthesized one, respectively, and $\overline{\mathbf{c}^n}$, $\overline{\hat{\mathbf{c}}^n}$ correspond to the estimated mean values of the n_{th} cepstrum vector in the real and the synthesized sets. N_r is the number of real speech frames, N_s is the quantity of codebook entries, and L is the number of cepstral coefficients used for the spectrum representation.

4.2. Testing the contribution of formants on acoustic-to-articulatory mapping systems

Evidence from using an articulatory synthesizer The articulatory synthesizer used in present section is explained in section 3.2. The Maeda's parameters are shown in figure **3-4**. The articulatory entries are the articulatory parameters in Figure **3-4**. In which each articulatory configuration generates a particular group of values of formants; therefore making the acoustic counterpart. The codebook is designed in such way that the distribution pattern in two dimensions F_1-F_2 covers a wide range. Scatter plots showing the relation between articulatory parameters and formants are obtained.

Evidence from human articulatory data In order to estimate the relevance of the MFCC coefficients and formants we have created a measure which consist of taking the average values of mutual information out of the entries characteristics. That is, average MI values

for formants is calculated by using the expression,

$$\bar{\mathcal{I}}(\boldsymbol{F}, y^m(t)) = \sum_{i=1}^{4} \mathcal{I}(\mathrm{F}_i, y^m(t)) \tag{4-9}$$

where the F_i are the formants and $\boldsymbol{F} = [\mathrm{F}_1 \cdots \mathrm{F}_4]^T$ and y^m are the $m = 1, ..., 14$ EMA channels.

The average of MI values for the MFCC coefficinets is calculated as follows,

$$\bar{\mathcal{I}}(\boldsymbol{\zeta}, y^m(t)) = \sum_{i=1}^{13} \mathcal{I}(\zeta_i, y^m(t)) \tag{4-10}$$

where $\boldsymbol{\zeta}$ is the vector of MFCC values. The number 13 corresponds of the quantity of the MFCC coefficients, which is commonly used in speech processing procedures.

Testing the contribution of formants in a system based on artificial neuronal networks
This section is intended to show the usefulness of the formants in acoustic-to-articulatory mapping systems based on artificial neural networks. MFCC coefficients are widely used in speech signal processing. To represent the speech signal, two input sets are created, the first one is formed by the MFCC coefficients, and the second uses the MFCC coefficients with the formants. The performance improvement can be used as an indication of the amount of relevant information that is added by the formants. The parameters are estimated in sync with the articulatory signal. Three layers form the neural network. The second layer has tansig activation functions; and, the output layer is a neuron with linear activation function.

Training and testing sets are constructed for the speaker *fsew0* from the `MOCHA-TIMIT` database. The training set is formed of 368 files and the testing set is composed of 46 files. The scaled conjugate gradient (SCG) optimization algorithm was used for the training process. First–order optimization algorithms, such as the standard backpropagation gradient descent, use the first derivative value of the error function; in contrast, conjugate gradient optimization algorithms use the second derivatives of the error function, which makes them considerably faster than the standard backpropagation algorithm [**?**]. The requirement of memory of the SCG optimization algorithm is less than other second–order methods. An additional advantage is that SCG does not require the user to look for important tuning parameters like the rate of learning. For these reasons, we selected SCG as the method to estimate the parameters of the neural networks.

EMA channel	Articulator
li_x-li_y	lower incisor
ul_x-ul_y	upper lip
ll_x-ll_y	lower lip
tt_x-tt_y	tongue tip
tb_x-tb_y	tongue body
td_x-td_y	tongue dorsum
vl_x-vl_y	velum

Table 4-1: EMA channel names. The suffix $_x$ denotes the x-coordinate of the coil and suffix $_y$ corresponds to y-coordinate.

4.3. Testing the relevant maps of TF features

4.3.1. Articulatory data

The present study uses the `MOCHA-TIMIT`, which made use of an AG-100 EMA system. It includes the acoustic-articulatory data of two speakers. One is female (fsew0), and the other is male (msak0). The sentences in this database are designed to provide phonetically diverse material in order to maximize the usefulness of the data for speech technology and speech science research purposes [89]. It is composed of 460 short phrases. The `MOCHA-TIMIT` database includes four data streams recorded concurrently:

- Acoustic waveform recorded at 16 kHz sample rate and with 16 bit precision.
- Electromagnetic articulograph waveforms contaning the positions of ten coils in the midsagittal plane. Movements of receiver coils attached to the articulators are sampled by the EMA system at 500 Hz.
- Electropalatohgraphy patterns wich correspond to 62 binary values provided by 62 contact attached to hard palate.
- Laryngograph signal recorded at 16 kHz sample rate.

Coils are affixed to the lower incisor (li), upper lip (ul), lower lip(ll), tongue tip (tt), tongue body (tb), tongue dorsum (td), and velum (vl). The two coils at the bridge of the nose and upper incisors, respectively, provide reference points to correct errors produced by head movements. See Figure **4-1**.

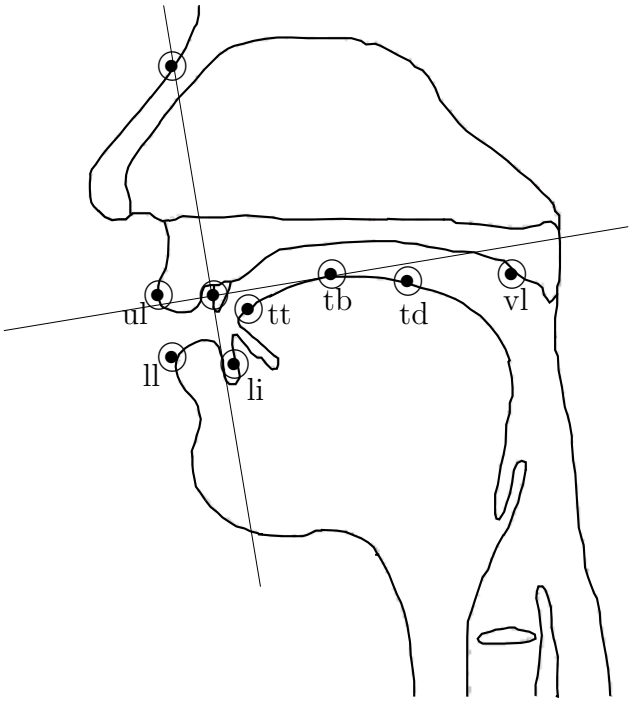


Figure 4-1: Positions of EMA contacts in the MOCHA-TIMIT database. tt, tongue tip; tb, tongue body; td, tongue dorsum ; li, lower incisors; ll, lower lip; ul, upper lip; vl, velum.

Data preprocessing

Label files of `MOCHA-TIMIT` database are used to discard silent segments at the beginning and the end of the utterances. The EMA trajectories are resampled from 500 Hz to 100 Hz after a filtering process with an 8^{th} order Chebyshev Type I low-pass filter of 40 Hz cut-off frequency. Since muscle contractions typically have bandwidths of up to 15 Hz [40], the 40 Hz value is apt to retain the information corresponding to articulators movement. The filtering process of EMA signals was carried out in both forward and reverse to remove phase distortions.

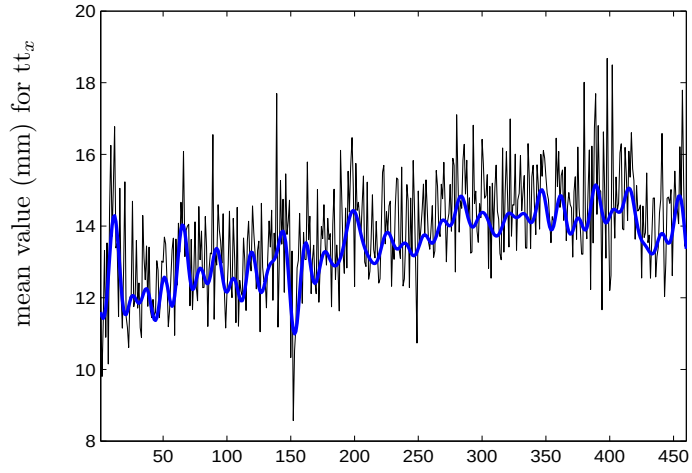
A process of standardization as the suggested in [89, 90] is done. The conventional process of standardization calculates the average values and the global standard deviations and then they are applied in the EMA pathways , but this may cause difficulties due to the change on average values from one phrase to another in the recording process. While the rapid changes of the average values are given for the phonetic content in each phrase, the slow changes are mainly caused by the articulatory adaptation of the subject during the recording session. It is useful to eliminate the second type of variation while keeping the other one.

This is carried out by subtracting one version of the average values obtained when moving the vector of average values, whose dimension is 460, through a low-pass filter. The value is fixed heuristically such that 15 % of bandwidth is low-pass filtered. The vector of values to be substracted is shown in blue in figure **4-2**. This process is similar to using a moving average mean, instead of using a global mean across all utterances. Once the low-frequency trend of means is removed from the articulatory data, a standard normalization is carried out.

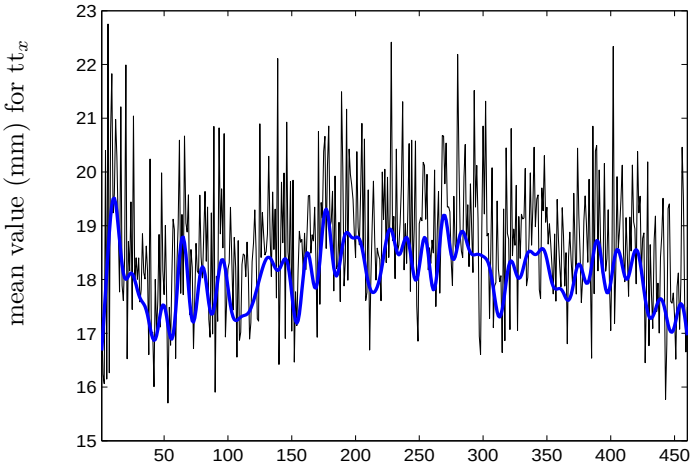
4.3.2. TF relevant features for inversion

For the sake of constructing the maps of relevant features, the statistical measure of association is applied to the time–frequency atoms enclosed in the context window $[t-t_a, t+t_b]$, where $t_a = 200$ ms and $t_b = 300$ ms. A total of 50 frames taken every 10 ms in time are parameterized using the 24 wavelet packet filter banks, as described in section §2.2.3. The process generated 1200 statistical association outcomes for each time t. The maps are constructed using 10 ms shift rate, the same used in [90, 110, 122, 6]. It is worthy to mention a recently developed method based also on GMM's having better performance over MOCHA-TIMIT database, which instead of using 10 ms time–shift, a frame step of 18 ms is utilized [77]. However, we selected 10 ms shift rate because it is commonly used in the literature.

For testing the usefulness of the relevant maps GMM based regression is carried out. Relevant features are tested on both speakers and the average RMSE and average correlation are measured along both speakers. The number of inputs is varied ranging from $p = 24$ to $p = 168$ ($p = 24, 72, 120$ and 168); that is, 1, 3, 5 and 7 frames around current time of analysis are taken into account. The input vector is transformed using Principal Component Analysis, where $n_p = 24, 35, 35, 50$ components are taken, respectively. The number of principal components was selected such that information retained holds at least between 92 % and 93 %. In the case of relevant maps, the $p = 24, 72, 120$ and 168 most relevant atoms are used. Then, the $n_p = 24, 35, 35, 50$ principal components are extracted to form the input vector. In all cases 32 mixtures are used. The model parameters are found by using the expectation maximization (EM) algorithm [11]. It must be quoted that the articulatory estimations are not low-pass filtered in this work. To measure the accuracy of the mapping a 5-fold cross-validation testing is carried out. The 460 sentences are divided into 5 partitions consisting of 92 sentences, and then one of the partitions is reserved for testing by turns, while the other 4 partitions are used for training, as discussed in [110]. The performance is measured by using the root mean square error and the Pearson's correlation coefficient.



(a) female spaker



(b) male spaker

Figure 4-2: A plot of mean tongue tip x coordinate calculated for each utterance from the MOCHA database for female speaker in a) and for male speaker in b). In addition, it is shown in blue the underlying trend captured by low-pass filtering the vector of raw means. This trend signal correspond to the mean values used in the normalization process.

Three experiments are performed: first, the Kendall coefficients are estimated over the whole speech samples; second, the Kendall coefficients are obtained for plosive consonants; and third, the relevant maps are obtained over phones for which a given articulator is critical. In addition, the TF relevant are tested on the subject-independent acoustic-to-articulatory inversion of fricative sounds, but mel filter bank is used instead of WP-based filter bank.

When one articulator constricts for a phoneme, the others are relatively free to coarticulate (if they do not cause an additional constriction). Because non-critical articulators are free to move, the statistical association measure could be affected by the intrinsic movements of these articulators. Furthermore, non-critical articulators could not been affecting notoriously on the acoustics of the speech signal. Therefore, relevant maps are also estimated in the framework of critical articulators.

The correspondence between articulators described by EMA pellets and their role in production of phonemes, critical or non-critical, is obtained in [43], as follows:

- ul_y : /p, b, m/
- ll_x : /f, v/
- ll_y : /p, b, m, f, v/
- tt_x : /θ, ð, s, z, ʃ, ʒ, ʧ, ʤ/
- tt_y : /θ, ð, s, z, ʃ, ʒ, ʧ, ʤ, t, d, n/
- td_y: /k, g, ŋ/
- v_x: /m, n, ŋ/

4.3.3. Measures of performance

Commonly, root mean square error and linear correlation have been used for the assessment of acoustic-to-articulatory inversion procedures [80, 40, 104, 90, 39, 110, 65, 77]. RMS error is an indication of the overall distance between two trajectories. The correlation score is an indication of similarity of shape and synchrony of two trajectories. The RMS error is calculated separately for each articulator as follows,

$$E_{RMS} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y^m(t) - \hat{y}^m(t))} \tag{4-11}$$

where N is the number of input-output vector pairs, or patterns, in the test set, $\hat{y}^m(t)$ is the value estimated by the m_{th} neural network, and $y^m(t)$ is the measured or true value. Besides it is calculated the correlation value between the EMA trajectories and the estimated trajectories by using following expression,

$$\rho = \frac{(y^m - \bar{y}^m)(\hat{y}^m - \bar{\hat{y}}^m)}{\sqrt{\sum_t (y^m - \bar{y}^m)^2 \sum_t (\hat{y}^m - \bar{\hat{y}}^m)^2}} \tag{4-12}$$

where $\bar{y}^m$ and $\bar{\hat{y}}^m$ are the means of y^m and $\hat{y}^m$, respectively.

5 Role of formants on articulatory inversion

5.1. Resulting cepstral distance

Recovering articulatory trajectories in the context of acoustic-to-articulatory consists of choosing every time a single articulatory vector among several candidates retrieved from the codebook [76]. Lifter optimization is carried out using an evaluation strategy that preselects ten candidate entries, out of which only the one with the best JND score is taken into account for the cost function. That is, this cost function does not assign errors when at least one entry among ten generates a synthesized speech signal whose formant frequencies are sufficiently close to the real ones. This strategy is in agreement with the procedures commonly used in acoustic-to-articulatory inversion by analysis-by-synthesis approaches.

The first stage of the optimization process is the construction of a grid with respect to the parameters s and T of Eq. 3-63, which allows spotting regions where local minima are located. The selection of the step size for both parameters is made using a tradeoff between preciseness and computational cost. Too wide steps do not permit a reliable identification of local minima regions; on the other hand too narrow steps increase considerably the computational load. It was heuristically found that using a step size of less than 0,28 for the parameter s does not produce considerable changes to the expanded region where local minima are located. Similarly, a step of size 2 for the parameter T produces small changes in the lifter shape. Thus, step sizes of 0,28 and 2 for parameters s and T, respectively, were selected for the grid construction.

Next, we select the region to be explored. This region should contain the significant local minima of the cost function. Using $s = 1$ and $T = 60$ in Eq. (3-63) we obtain a shape similar to the linear lifter. Values of $T > 60$ produce quadratic lifters, which is a form that has a small probability of success since it would enhance the higher quefrency components.

	constant	Linear	Juang	Meyer	B
10 candidates	19.95	14.54	19.31	16.92	**12.00**

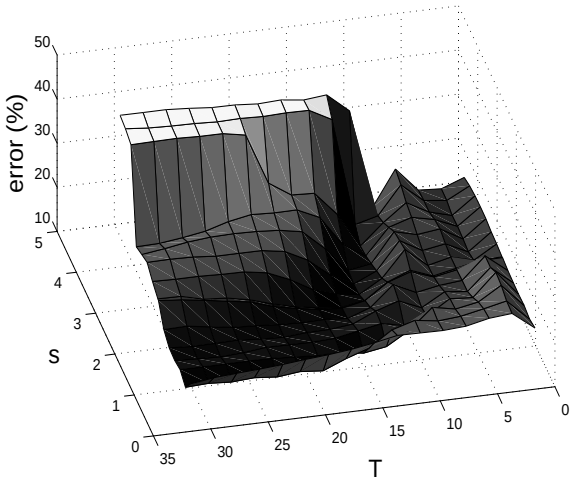
Table 5-1: Percentage of speech frames with formants outside the just noticeable difference (JND) in codebook formed by cepstral coefficients.

The rectangular region enclosed by the points $s = 0{,}44$, $T = 31$ and $s = 3{,}5$, $T = 61$ was also explored using larger step values for s and T, however no relevant minima were found. Therefore, the region included in the optimization process was the rectangle enclosed by the points $s = 0{,}44$, $T = 1$ and $s = 3{,}5$, $T = 31$. The 3D and contour plots of the this cost function are shown in Fig. **5-1**. The optimization method is carried out in the neighborhood of the local minima. After optimization process the result is $s = 2{,}66$, $T = 13{,}35$, which we label B. The value of the cost function at B is 12,1 %. The corresponding lifter shape is shown in Fig. **5-2**.

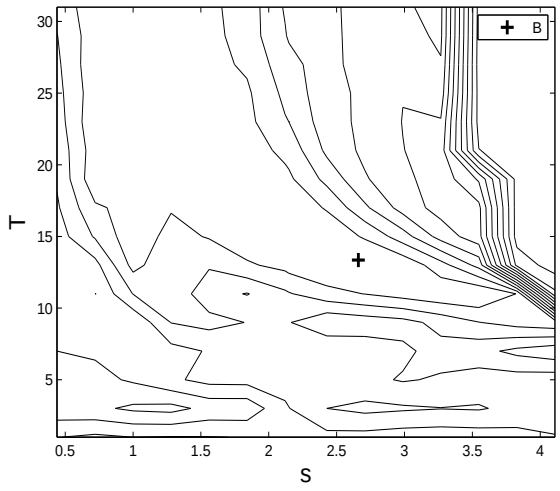
Evaluation of lifters Table **5-1** shows JND results for the cost function already described. It is evaluated for the commonly used constant, linear, Juang and Meyer lifters, as well as for the lifter B. For the case of ten candidates, there is respectively an improvement of 17.2 % of lifter B compared to the linear lifter.

We assess the performance of the lifters commonly found in speech recognition applications when using cost functions involving more than ten candidates. These results are summarized in Fig. (3) which shows the JND scores for Juang, Meyer, and B lifters, for up to 100 candidate entries used in the cost function.

Application of the optimized lifter (B) modifies the relationship between formant and cepstral distances. The modification of the pattern of Fig. **1-5** is shown in Fig. **5-8**. The proposed measure has the effect that cepstral distance relates better with formant distance; that is, minimal cepstral distance between real speech and codebook entries implies minimal formant distance.



(a)



(b)

Figure 5-1: 3D and contour plots of the ten candidate cost function with respect to the parameters s and T in Eq. (3-63).

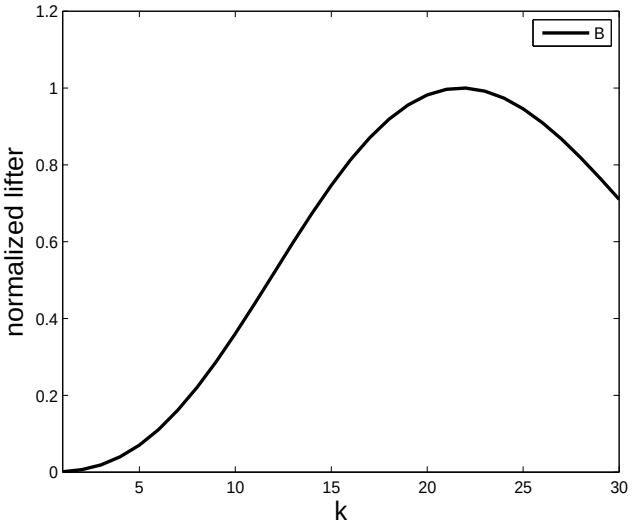


Figure 5-2: Lifter shape for $s = 2{,}66$, $T = 13{,}35$ (point B).

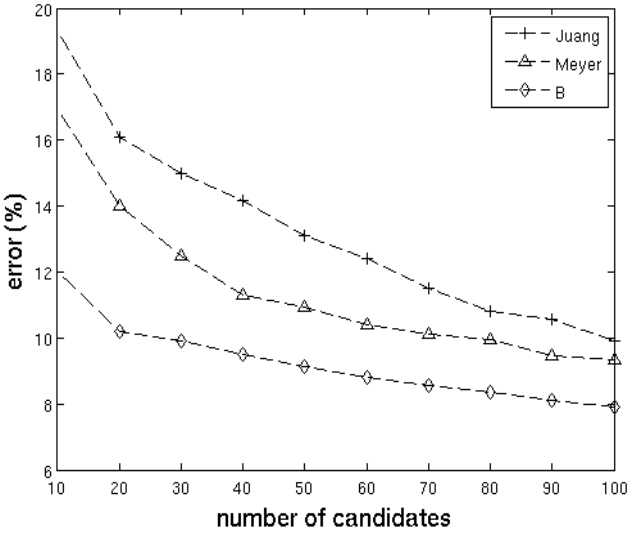


Figure 5-3: JND error (%) of lifters with respect to number of selected candidate codebook entries.

5.2. Contribution of formants on inversion systems based on neural network

Evidence from using an articulatory synthesizer The dispersion diagrams between the formants and some of the articulatory parameters are shown in Fig. **5-4**. It can be observed i Figure **5-4**(a) that the value of the first-formant (F_1) tends to increase with the value of the tongue dorsum (Maeda's parameter P_2). However, it is also observed that the F_1 value not only depends on P_2, due to the fact that many configurations can generate visible changes in F_1. In Figure **5-4**(c) it can be observed that F_3 decreases as the position of the articulatory parameter P_4 increases. Similarly, many articulatory configuration can produce enormous changes on the formant F_3. Figure **5-4**(b) shows the dispersion diagram of the second formant versus P_3 (tongue dorsum shape).

Results from using artculatory data The Figure (**5-5**) shows the $\bar{\mathcal{I}}(\boldsymbol{F}, y^m(t))$ and the $\bar{\mathcal{I}}(\boldsymbol{\zeta}, y^m(t))$ values for the 14 EMA channels, whose estimation is explained in equations (4-9) and (4-10).

Results from a system based on artificial neural networks This section shows the improvement in performance as measured by the Root Mean Square error (RMSE) and the Pearson when formants are added to the MFCC vector for each of the articulatory trajectories. Acoustic-to-articulatory inversion experiments using MFCC and MFCC$\oplus$formants are carried out.

Results can be observed in Figure (**5-6**). The RMSE value and the correlation are calculated for the case in which 6, 10, 14, 18 and 22 neurons are used in the intermediate layer. It can be observed that the results for the entry set MFCC$\oplus$formants are better than for the case in which only MFCC coefficients are used. Even when modifying the quantity of parameters in the neuronal network the results keep stable.

The improvement in the performance for each of the channels when 14 neurons are used in the intermediate layer is shown in Figure **5-7**. When Figure **5-7** is compared with Figure **5-5**, a similarity can be observed. Particularly, higher statistical association values which imply better performance of RMSE and correlation. For the case of 14 neurons in the hidden layer and including the formants to input feature set, the performance is improved in terms of RMSE in 2.5 % and 2.9 % for the case of the correlation measure. These results

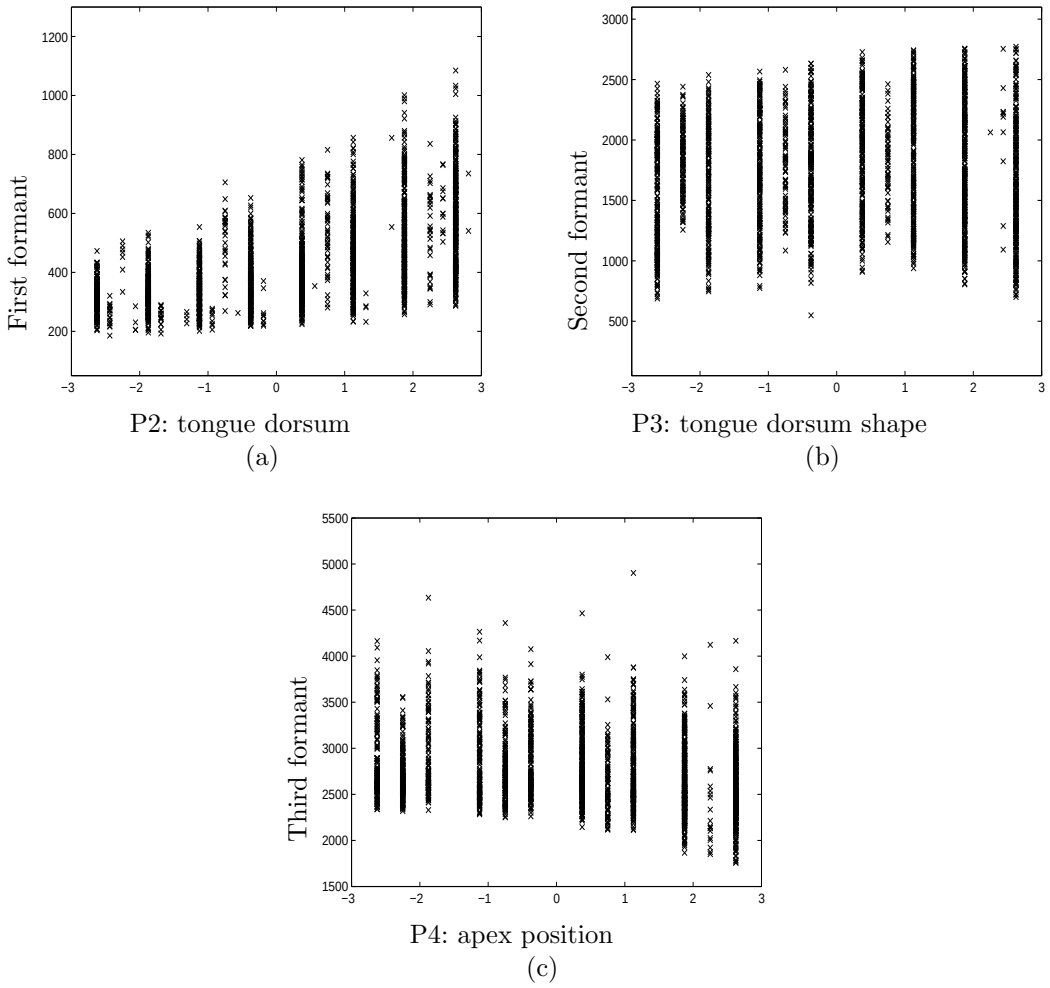


Figure 5-4: Scatter plots of formants versus articulatory parameters : a) first formant vs. tongue dorsum position (P2), b) second formant vs. tongue dorsum shape (P3) and, c) third formant vs. apex position (P4)

are in agreement with those ones shown in [77].

5.3. Discussion

One of the ways to analyze the relation between the movement of the articulators and the formants is by using articulatory synthesizers [70]. However, it is preferable to do this analysis by using real data. This allows us to make more reliable analyses. In this chapter, the advantage offered by the EMA technology was used in order to analyze statistically the relation between the formants and the position of the articulators.

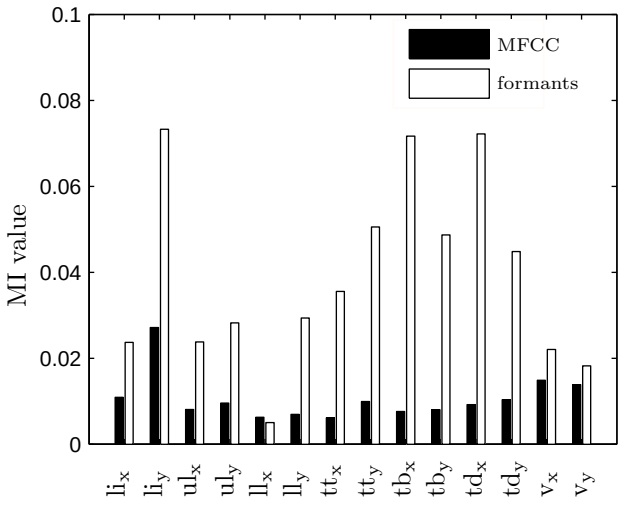


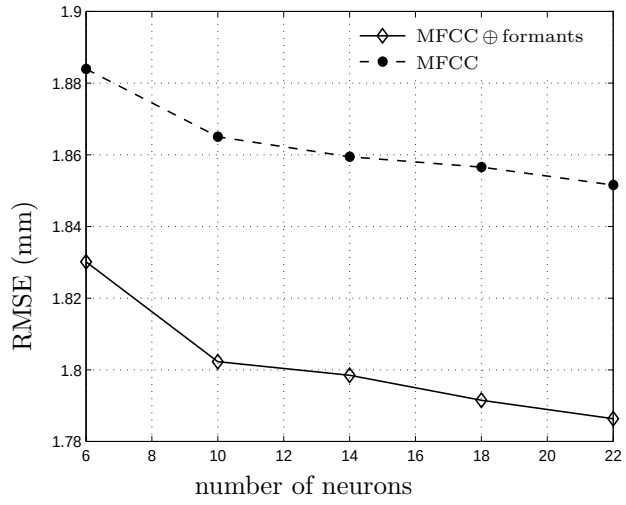
Figure 5-5: Measures of mutual information between the position of the two articulators and the different type of entries, MFCC $\bar{\mathcal{I}}(\boldsymbol{\zeta}, y^m(t))$ (in black) and formants $\bar{\mathcal{I}}(\boldsymbol{F}, y^m(t))$ (in white)

In [77] it is shown that when adding formants to an articulatory inversion system based on GMM the error rate decreases 3,4 % and the correlation increases 2,7 %. In this chapter, which is based on artificial neuronal networks, the error is decreased 2,2 % approximately and the correlation increases 2,8 %, which are comparable with [77]. On the other hand, the mutual information measure allows us to observe that the formats are highly influenced by the tongue, the main articulatory organ. The regression experiment based on neuronal networks of this paper allows us to confirm this finding due to the fact that when adding the formants, its positive effect is especially seen in the inference of the movement the apex, the body and the back of the tongue.

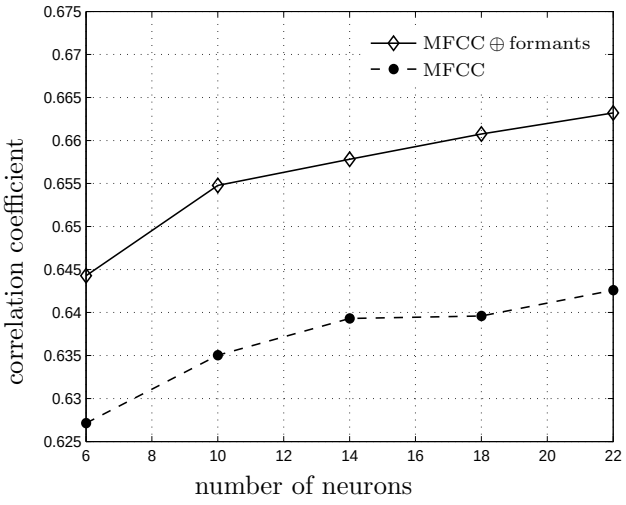
From the statistical point of view it is shown that there is an intrinsic relation between the formants and the position of the tongue (see Figure **5-5**) which provokes an improvement in the performance of articulatory inversion systems as it is also shown (see Fig **5-6**). This finding allows us to establish that the articulatory inversion systems found in the state of the art would have a better performance if they included formants within the group of parameters representing the speech signal.

The derived lifter by the optimization process performed in this work is shown to outperform other lifters previously presented in the literature, see figure **5-3**. The initial goal of having a liftered cepstral distance that suppresses sufficiently glottal variability, so that a minimal distance in the cepstral domain implies a minimal distance in the formant

domain is almost achieved, as can be observed in Figs. **1-5** and **5-8**.



(a)



(b)

Figure 5-6: Measures of performance using the set of entries MFCC and MFCC⊕formants in terms of RMSE (a.) and the Pearson correlation (b.) for several numbers of elements in the intermediate layer of the neural network.

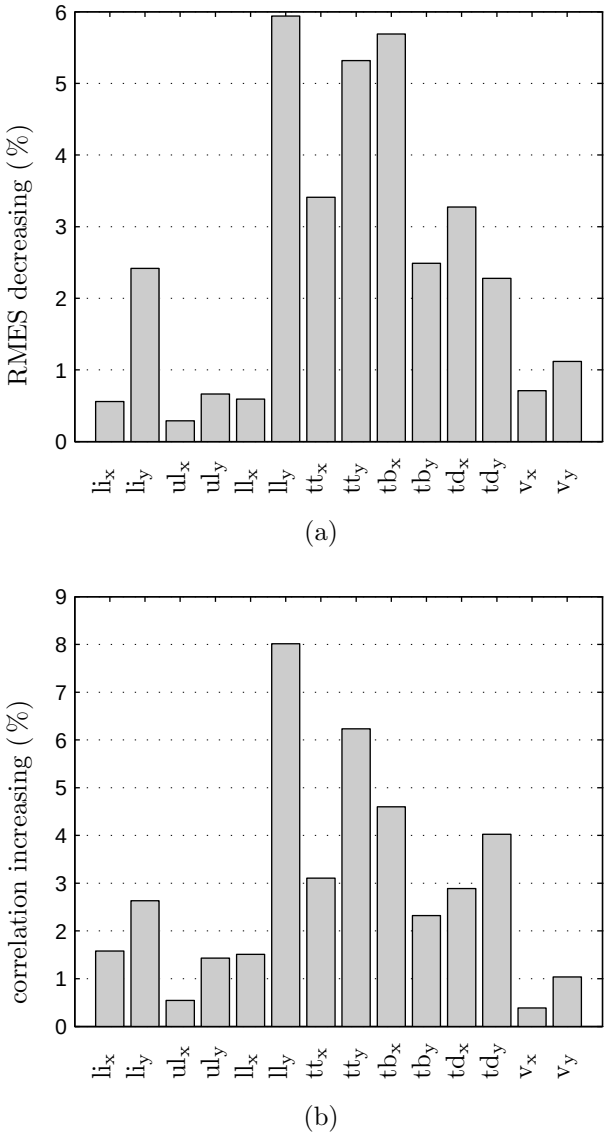


Figure 5-7: Performance improvement in relation to RMSE and the correlation to different positions of the articulators. a) Decreasing error percentage; b) Increasing correlation percentage

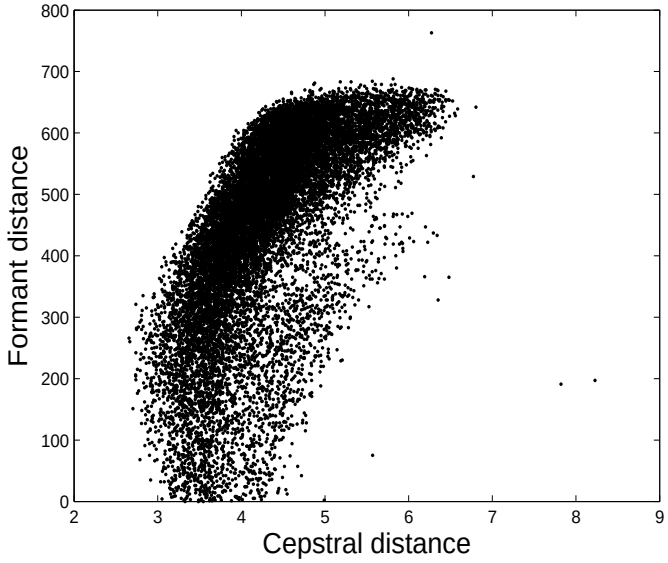


Figure 5-8: First formant vs. liftered (using lifter A2) cepstral distances from all entries in the articulatory-to-acoustic codebook to a real speech segment coresponding to an /a/ sound. In comparison with **1-5**, formant and cepstral disatnces are better related.)

6 Maps of relevant TF features as a mean for improving performance of the acoustic-to-articulatory mapping

6.1. On estimating time-frequency maps of relevant features

Relevant maps using whole speech signal From each one of the speaker sets, fsew0 and msak0 holding 98000 and 92000 elements, respectively, the number of 15000 pairs $\{\boldsymbol{X}_t,\, y^m(t)\}$ of EMA–acoustic information are randomly selected to estimate the relevant maps belonging to fsew0 and msak0. The statistical association between each variable $x(t+d, f_k)$ and vector $y^m(t)$ of the articulatory configuration $\boldsymbol{y}(t)$ is estimated. Calculation of each of the the 28 maps, 14 for each speaker, requires computation of 1200 coefficients. The estimated τ–coefficients in Eq. (3-3), are employed to get the maps based on the Kendall measure.

The following strategy is accomplished to obtain the level of significant association. The τ values between 50 random Gaussian vectors and each of the 14 EMA channels are estimated. Among the 14×50 values the one with maximal magnitude is selected. The value that results from this experiment is 0,026, which is approximated to 0,03. As a result, any Kendall association estimate that is less than 0,03 is considered as insignificant.

Relevant maps over stops A plosive consonant is produced by blocking the oral cavity at some point. The constriction can be formed by the lips, in case of bilabials /p, b/; the tongue tip for alveolar /t, d/; and tongue dorsum when producing velars /k, g/. Thus, these are the critical articulators for those sets of phonemes. For the construction of relevants maps corresponding to ll_y, tt_y and td_y at maximum of 5000 pairs $\{\boldsymbol{X}_t,\, y^m(t)\}$ of EMA–acoustic

points are taken. If the total number of samples is less than 6000, 2000 pairs are taken. The Kendall τ coefficient between each variable $x(t+d, f_k)$ and articulatory trajectories of the channels corresponding to ll_y, tt_y and td_y is estimated. The resulting points are used to construct the Kendall relevant maps. This procedure is performed for the female as well as for the male speakers in the MOCHA database. The maps are shown in Figure **6-1**. The zones of higher relevance are denoted by the brighter zones while the features that have relevance values less than 0,03 are black colored. As seen in all the graphs, in case of plosive phonemes the peak of maximal information is located after the current time of analysis $t = 0$.

Relevant maps of critical articulators In this section, the process used to achieve the relevant maps is carried out as in the case of plosive analysis. The Kendall τ coefficient between each variable $x(t+d, f_k)$ and articulatory trajectories of ul_y, ll_x, ll_y, tt_x, tt_y and td_y is obtained. At maximum of 5000 pairs $\{\boldsymbol{X}_t, y^m(t)\}$ of EMA–acoustic points for male and female speakers are randomly taken. Relevant maps are shown in Figures **6-2** and **6-3** for speakers fsew0 and msak0, respectively.

Some similarities in shape can be observed between the maps of female and male speakers, in particular for the case of upper lip y, tongue tip x, tongue tip y and tongue dorsum; however, the similarities are conditioned by frequency ranges. By contrast, in case of lower lip x and lower lip y,the position of brightest zones on relevant maps of female speaker are not in agreement with those positions on relevant maps of male speaker. We offer no explanations to this fact.

6.2. Assesment of Kendall coefficient for measuring the statistical dependence between articulators and speech components

$I^m_{d,k}$ values, given by Eq. (3-6), are used to accomplish the maps of relevant features based on χ^2 information measure. The information content estimates $\hat{I}^m_{dk}$ as well as the Kendall values τ^m_{dk}, for $d = -200, -190, \ldots, 290, 300$ ms and $k = 1, 2, \ldots, 24$, are used to form the relevance vectors $\boldsymbol{g}^m_I \in \mathbb{R}^{16800\times 1}$ and $\boldsymbol{g}^m_\tau \in \mathbb{R}^{16800\times 1}$, respectively. 16800 is the result from the product between 1200 atoms and 14 channels. The distribution of these two variables, $\boldsymbol{g}^m_\tau$ and $\boldsymbol{g}^m_I$, is shown in the scatter plot of Figure **6-4**. It can be seen the relationship between the two estimations, where most of the data can be explained by a

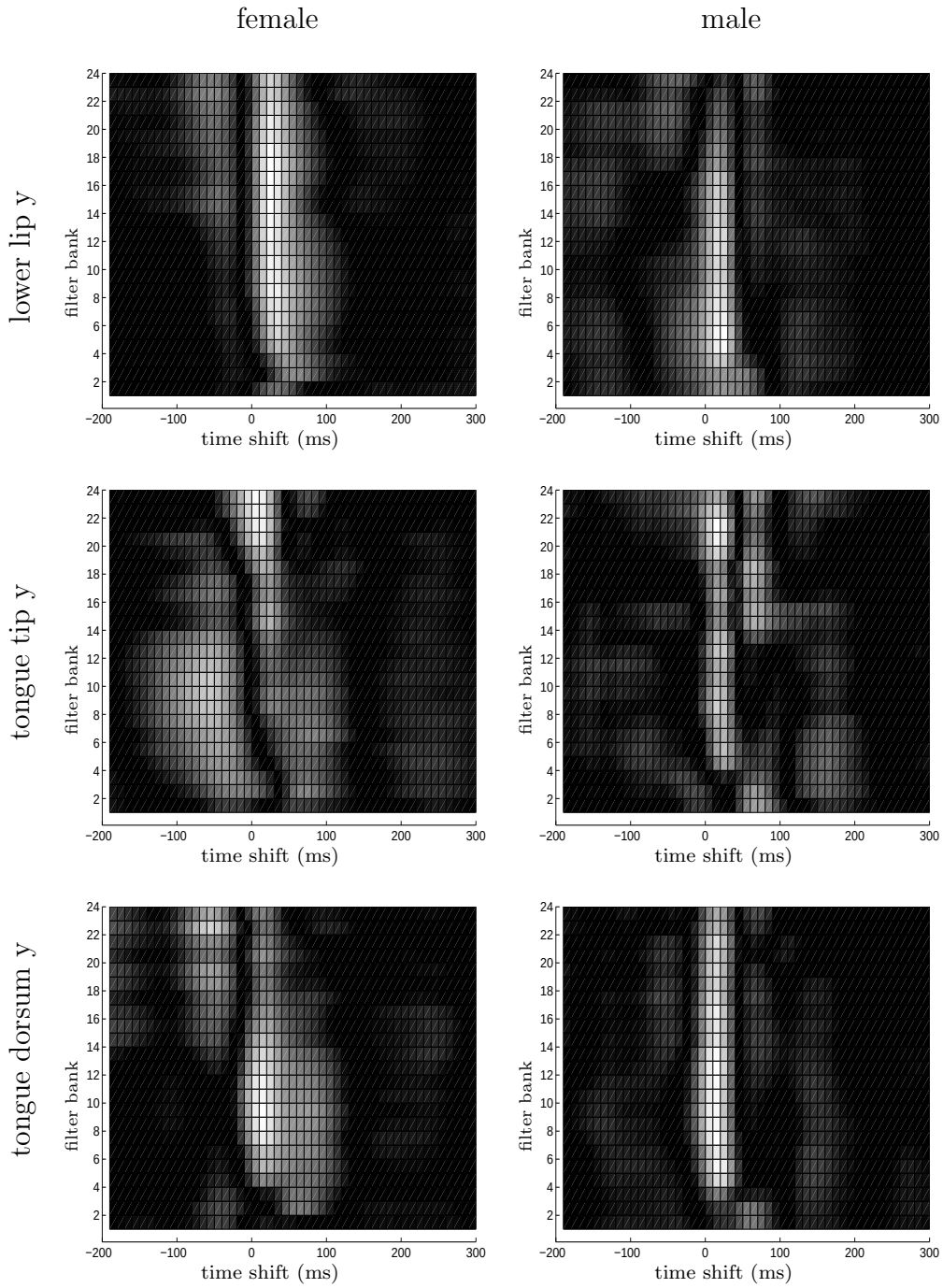


Figure 6-1: Relevant time–frequency atoms for the critical articulators of the stop consonants. Bilabial (/p, b/), lower lip y; alveolar (/t, d/), tongue tip y; and, velar (/k, g/), tongue dorsum y.

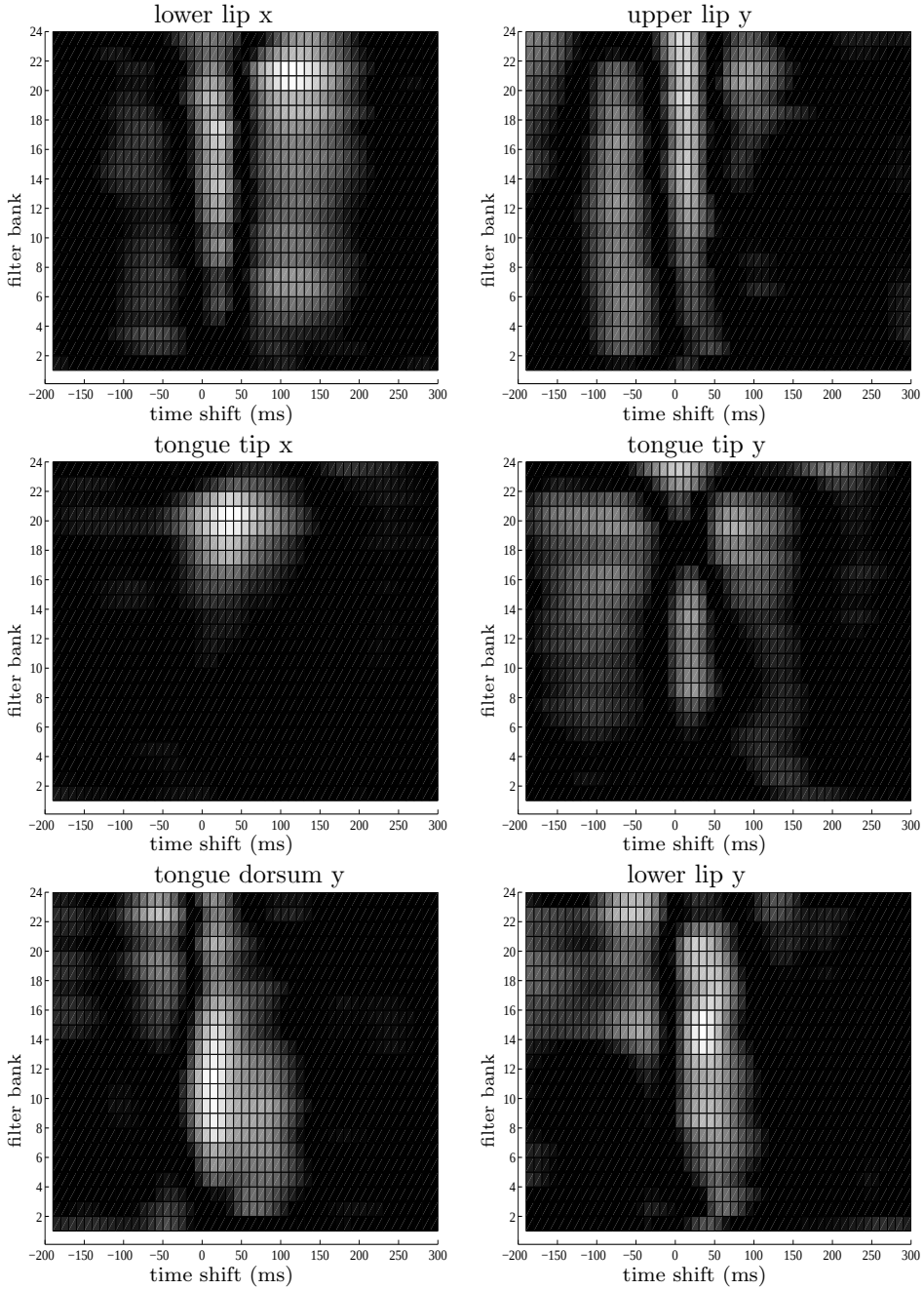


Figure 6-2: Relevant time–frequency atoms for female speaker in case of critical articulators.

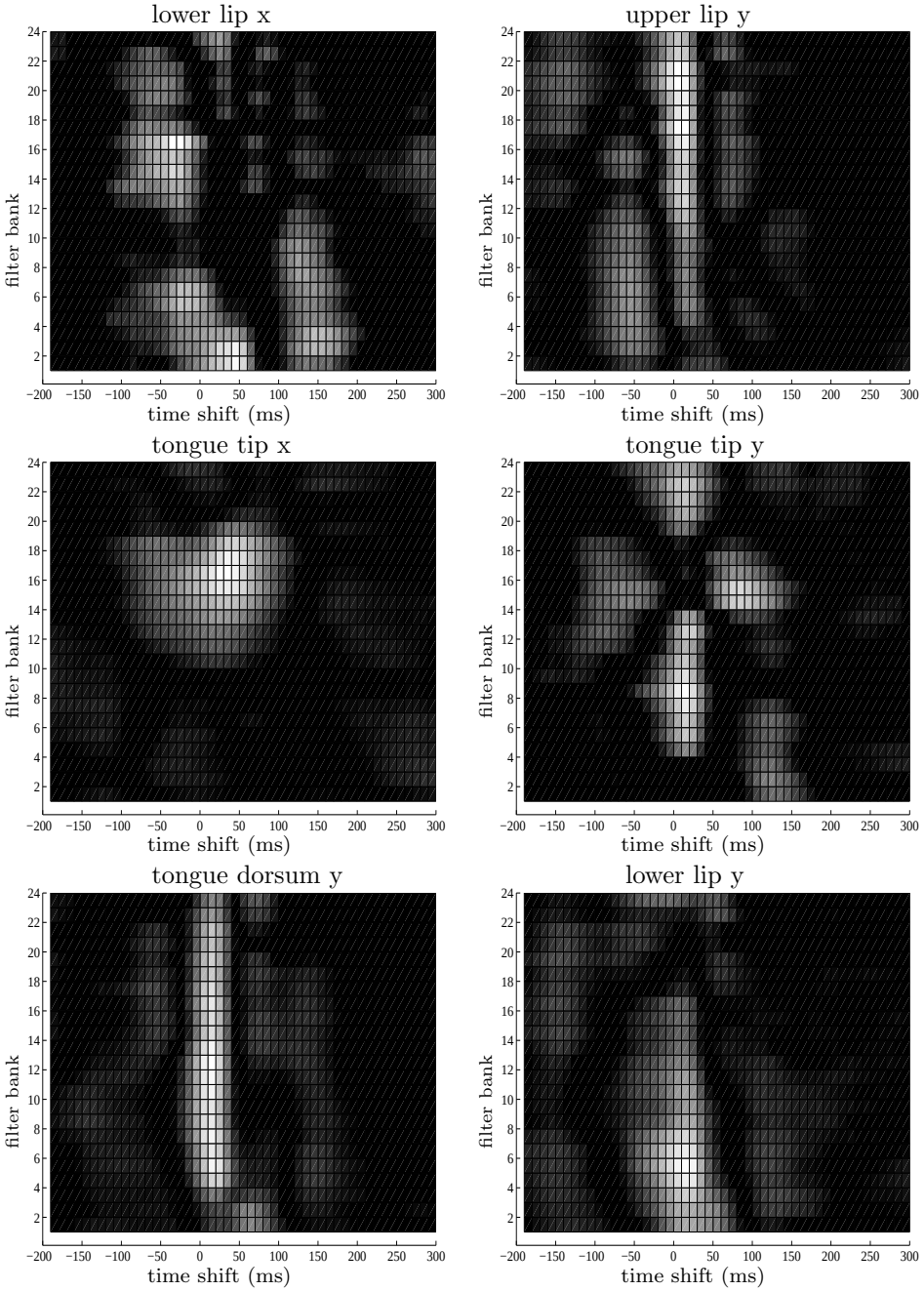


Figure 6-3: Relevant time–frequency atoms for male speaker in case of critical articulators.

second order fitting curve. In spite of outliers presence, the linear correlation value between the χ^2-information estimates and the τ^2 values is $\rho = 0{,}93$.

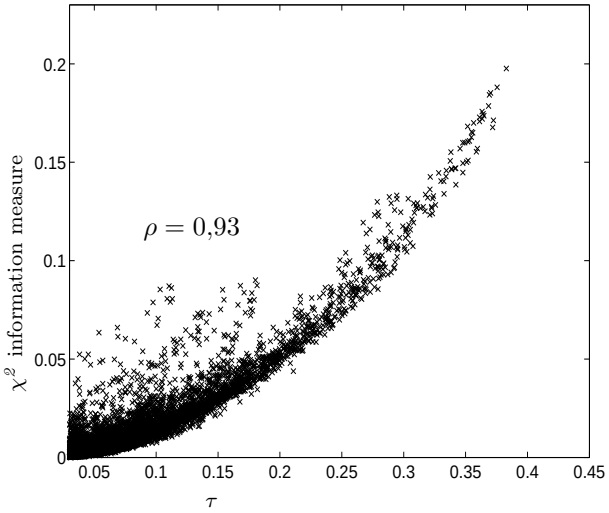


Figure 6-4: Scatter plots of the τ values versus the $\hat{I}$ estimates.

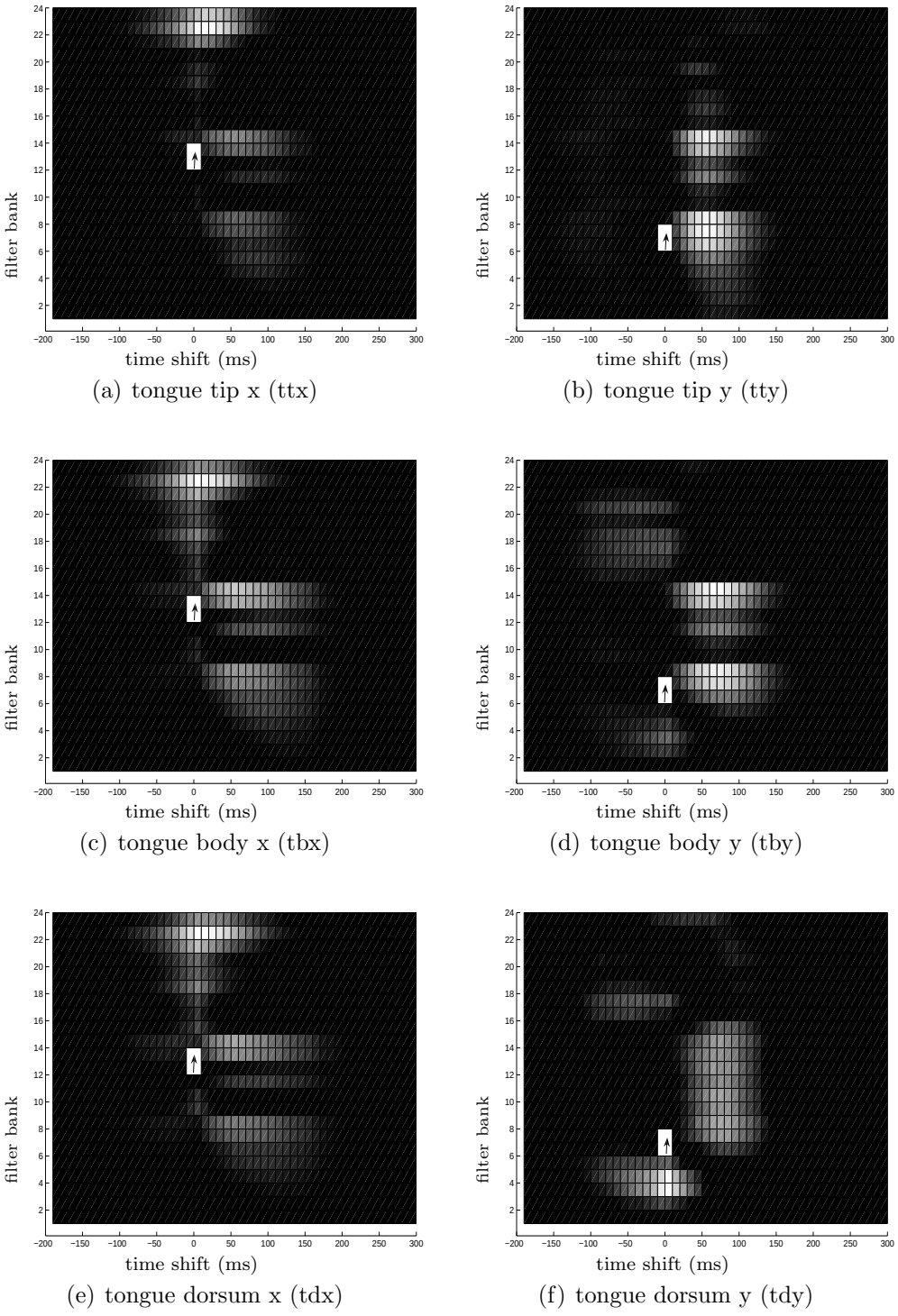
Information provided by inclusion of additional feature By using the partial Kendall correlation concept, the additional information provided by a new feature $x(t+d, f_k)$ located in the vicinity of the given feature $x' = x(t + d', f^m)$, denoted as $T(X(t + d, f_k), y^m(t); x')$, is estimated, where f^m is given by,

$$f^m = \underset{f_k}{\text{máx}} \ \tau(X(t, f_k), y^m(t)) \tag{6-1}$$

Estimation is carried out for each channel m, time–shift d and filter bank number k to attain the values $T^m_{d,k} = T(X(t + d, f_k), y^m(t); x')$ producing n_c conditional relevant maps.

Particularly, in case of ttx, tbx and tdx, f^m is fixed equal to 13 while for the dimensions tty, tby and tdy, $f^m = 7$. Figure **6-5** shows the distribution of the additional information provided by a feature, located at different frames as well as at different frequency bands, for the ttx, tty, tbx, tby, tdx, and tdy dimensions. The feature f^m is marked by arrow. It can be seen that the information added by a second feature $x(t + d, f_k)$, around the first feature x', is not symmetric. Besides, those features located after current position in time are more relevant.

Regression outcomes based on neural-networks Searching of most relevant sets of TF features regarding articulatory position inference is grounded on the following statement:



(a) tongue tip x (ttx)

(b) tongue tip y (tty)

(c) tongue body x (tbx)

(d) tongue body y (tby)

(e) tongue dorsum x (tdx)

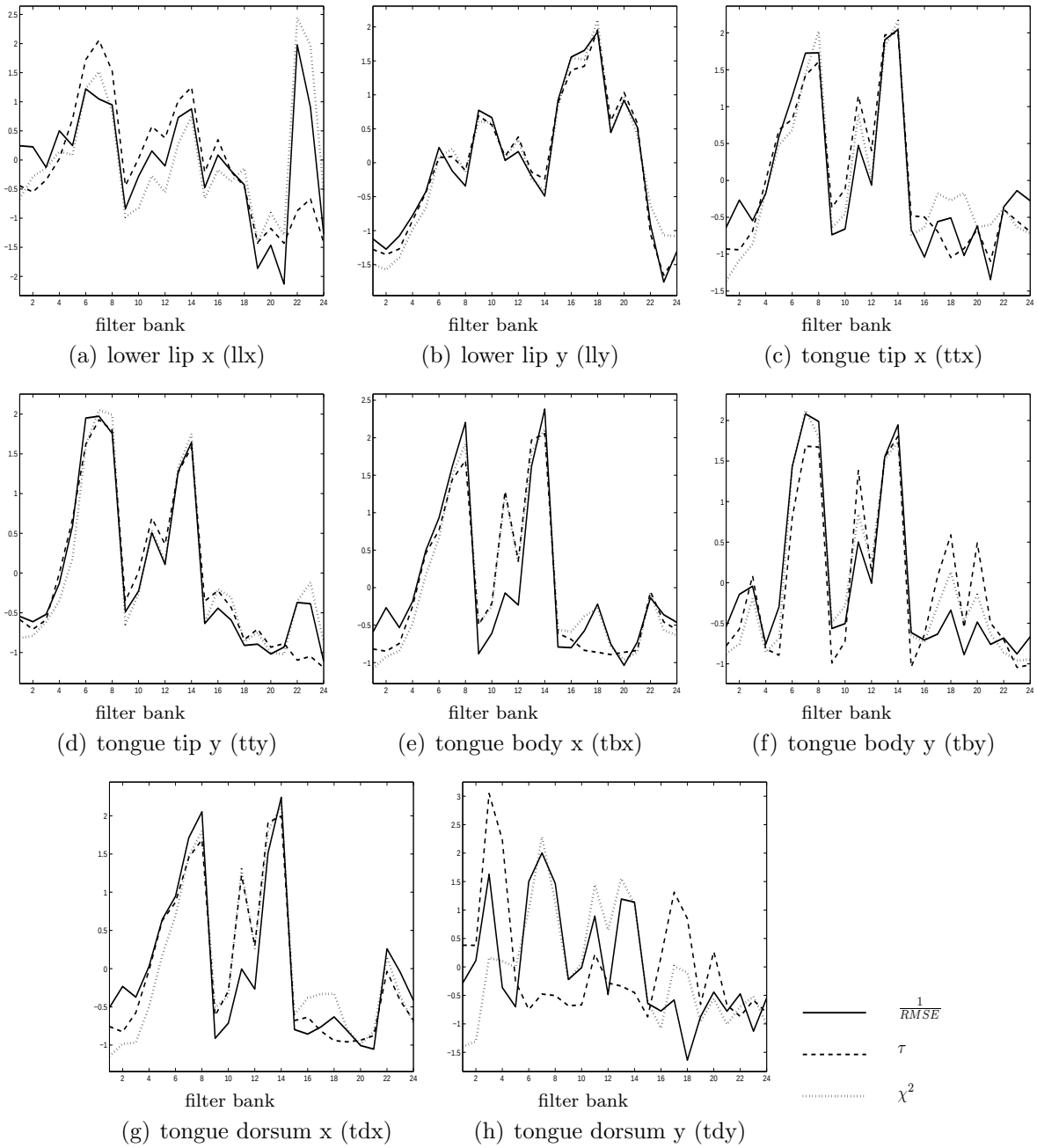
(f) tongue dorsum y (tdy)

Figure 6-5: Partial correlation when adding a new input feature $x(t+d, f_k)$ given the input denoted by the arrow in the picture. That is, the additional information provided by $x(t+d, f_k)$ given $x(0, 13)$ and $x(0, 7)$ in case of x–EMA–dimensions and y–EMA–dimensions of the tongue points, respectively.

the more relevant the features used by models - the better their performance. In this line of analysis, regression testing is carried out based on multilayer perceptrons (MLP), where the scaled conjugate gradient optimization algorithm is used during the training process. In particular, the training set is appraised by 79327 (80 %) input–output pairs while the testing set includes 19195 (20 %) elements. The neural models are trained during a number of 75 epochs that had been empirically fixed. Then, performance is calculated using the testing set. Performance of the neural models is determined as the inverse of the root mean square error value. Performed values are normalized for comparison with the statistical association and the partial correlation estimates.

In the beginning, validation of association between the estimated set of Kendall coefficients and the performance given by the neural network regression strategy is carried out using a single feature as the input to the regressor. In this case, each of the 24 filter banks located at time $t = 0$ (i.e., the position of the articulatory vector $\boldsymbol{y}$) becomes the input to each model. The hidden layer of the neural model has 10 neurons with *tansig* activation functions. At the same time, the output layer has a single neuron with linear activation function. Results of the model performance that are defined as the inverse of the root mean square error (RMSE) ($1/RMSE$) are shown in Figure **6-6**. Mostly, there is similarity between model performance and association values, but with exception of some bands in a few channels.

Later, two features are used as the inputs to the neural network to validate relationship between the measure of the information added by a new feature, which is quantified by using the partial Kendall coefficient (see Eq. (3-4), and the performance of a neural network regression system with these two features as the input set. In this instance, each model comprises 12 *tansig*-units in the hidden layer with a single linear activation function in the output neuron. The results of this experiment are compared with the partial Kendall correlation outcomes estimated for the frame located at time shift $d = 0$. In case of tbx, the pairs are formed by the filter $f_k = 13$ and the each of the other 23 filters. In case of tby, the filter $f_k = 7$ is used instead of the filter $f_k = 13$. Figure **6-7** illustrates similarities between the estimated partial correlation and the model performance (normalized inverse RMSE value).

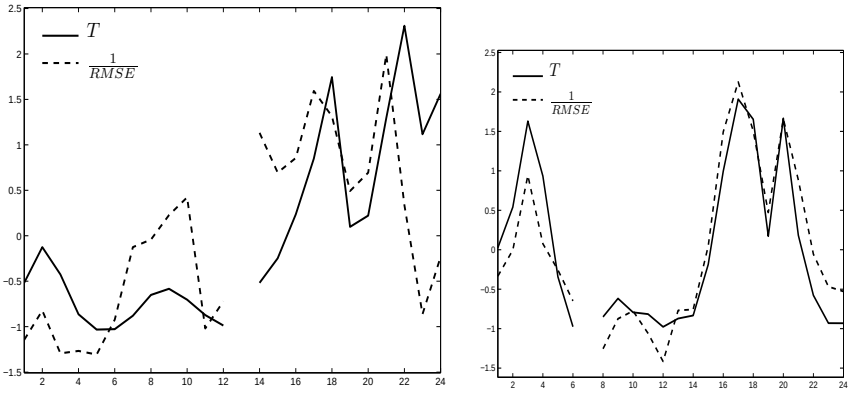


(a) lower lip x (llx) (b) lower lip y (lly) (c) tongue tip x (ttx)

(d) tongue tip y (tty) (e) tongue body x (tbx) (f) tongue body y (tby)

(g) tongue dorsum x (tdx) (h) tongue dorsum y (tdy)

Figure 6-6: Performance and statistical association estimations using the energy filter banks located at a time shift $d = 0$. Performance of the MLP regression strategy (noted as $\frac{1}{RMSE}$), nonlinear association estimates using Kendall τ coefficient(τ), nonlinear association values using χ^2 measure of information (χ^2).



(a) Partial correlation for the channel tbx

(b) Partial correlation for the channel tby

Figure 6-7: Comparison between partial T coefficient and neural network performance $1/RMSE$ in case of two input features. The partial correlation T is estimated between tbx (tby) and $x(t, f_k)$, given $x(t, 13)$ $(x(t, 7))$.

6.3. Relevant acoustic-to-articulatory maps using Gaussian mixture regression

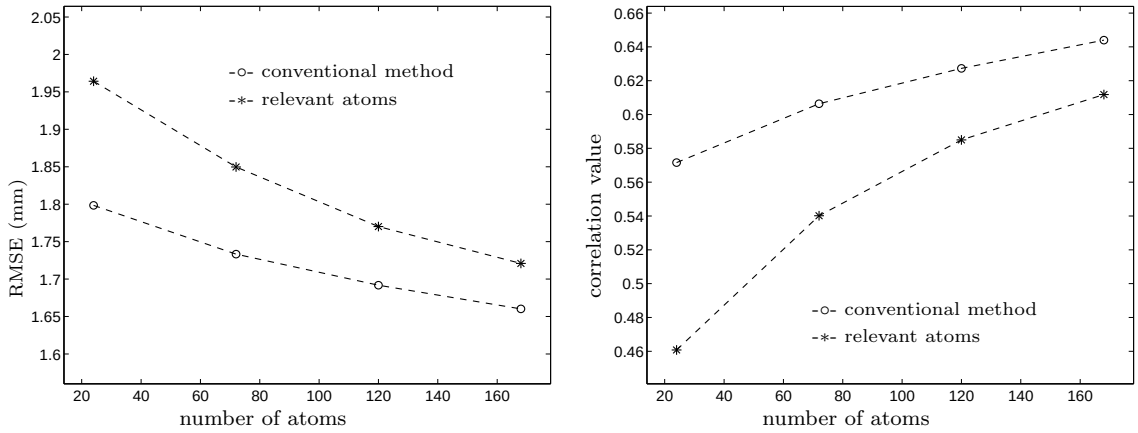
Significance test is performed in order to know whether the performance of proposed method is or is not significantly superior to conventional method's. The hypothesis H_0 (null hypothesis) and H_1 (alternative hypothesis) are compared, as in [77]. H_0 states that the performance of proposed method $(J(R))$ is not better than conventional method $(J(C))$. By contrast, H_1 states that $J(R)$ is greater than $J(C)$.

$$H_0 : d = J(R) - J(C) \leq 0 \qquad (6\text{-}2)$$
$$H_1 : d = J(R) - J(C) > 0$$

The hypothesis testing procedure is performed by using the Matlab command *ttest*, where the input argument is the vector formed by the 40 elements ($40 = 2$ speakers $\times 4$ input sets $\times$ 5 folds) corresponding to the performance outcomes for each articulatory channel.

Acoustic–to–articulatory mapping using the whole speech signal The process is performed for 24, 72, 120 and 168 input features giving the results shown in Figure **6-8**. Same Figure also depicts the results when using conventional method in case of taking 1, 3, 5 and

7 frames (24, 72, 120 and 168 input features) for the context window around current time of analysis. Full covariance matrix is used in the models.



(a) Average RMS Error value (b) Average correlation value

Figure 6-8: Average performance along both speakers of the conventional method and the method based on relevant atoms over the whole speech signal in terms of root mean square error (RMSE) and correlation measure.

Acoustic–to–articulatory mapping of plosives For each of the 5 partitions (consisting of 92 sentences) the phones corresponding to plosive phonemes are extracted and used to evaluate the relevant features obtained in section (6.1). One of the sets is reserved for testing by turns, while the other 4 sets are used for training. The process is performed using 24, 72, 120 and 168 inputs for both female and male speakers. For the sake of avoiding any possible problem caused by reduced number of samples available for training and testing processes, we choose diagonal co-variance matrix. The results, in terms of average RMSE and average correlation between both speakers, are shown in Figures **6-9** and **6-10**. It can be observed that the use of Kendall relevant maps improves the performance of the GMM based acoustic–to–articulatory regression systems for most of the selected quantity of atoms. In addition, we measure the average percentage of improvement along speakers for each of the selected number of atoms; and, these values are used to obtain the average improvement per articulatory channel shown in Table (**6-1**).

In case of conventional method, additional number of atoms (216, 312 and 408 atoms; 9, 13 and 17 frames) are used as inputs; and, the best performance among all selected number of frames, from 1 to 17, is taken. The value, termed ceiling, is depicted as dotted lines in Figure **6-10**. Some observations can be made : a) for ll_y, almost same performance is obtained

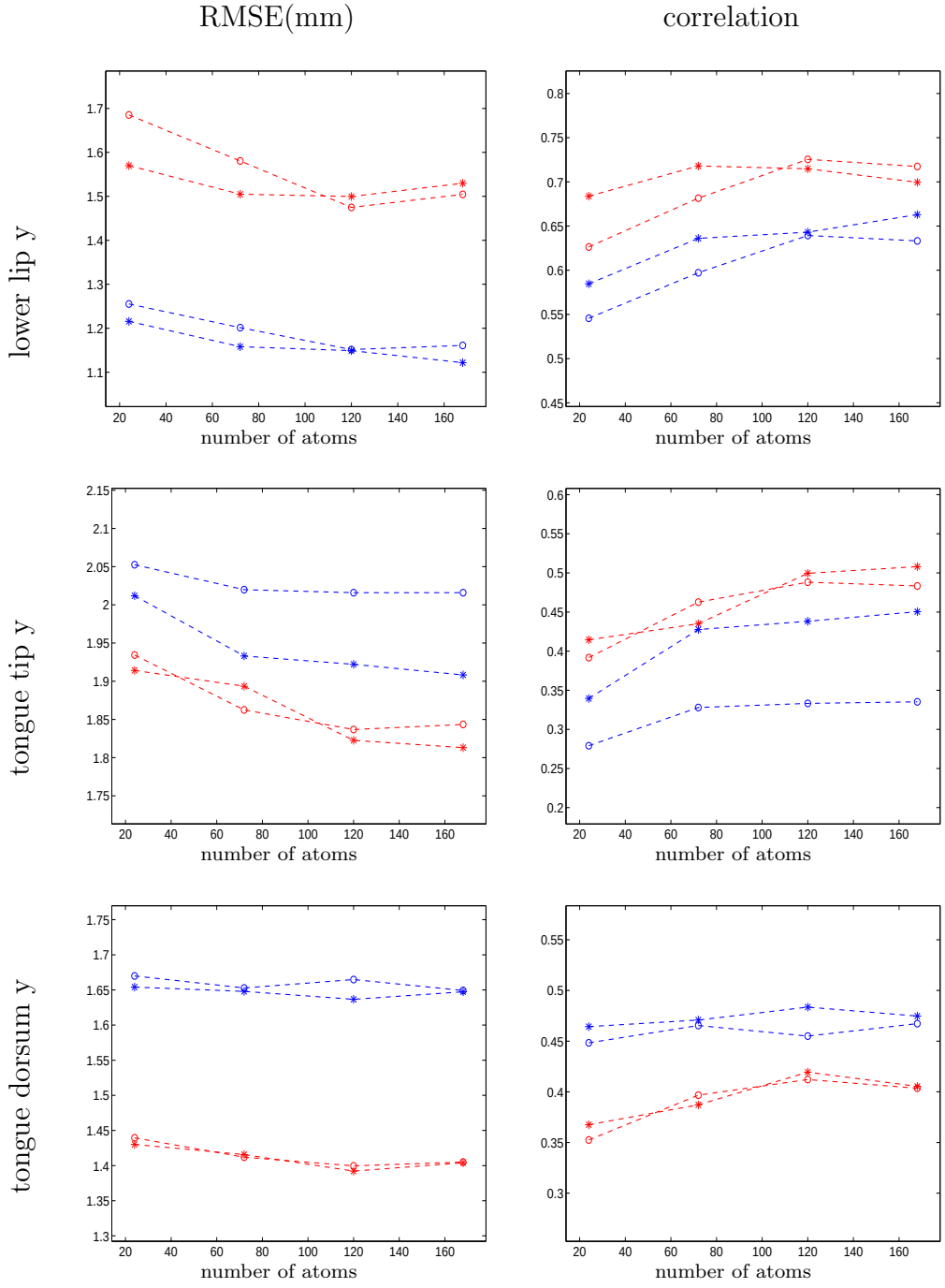


Figure 6-9: Performance in terms of RMSE and correlation using conventional method (noted with ∘) and using relevant time–frequency atoms (noted with ∗) for the critical articulators of the stop consonants. Bilabial (/p, b/), first column of figures; alveolar (/t, d/), second column; and, velar (/k, g/), third column. Results are colored in red and blue for female and male speakers, respectively.

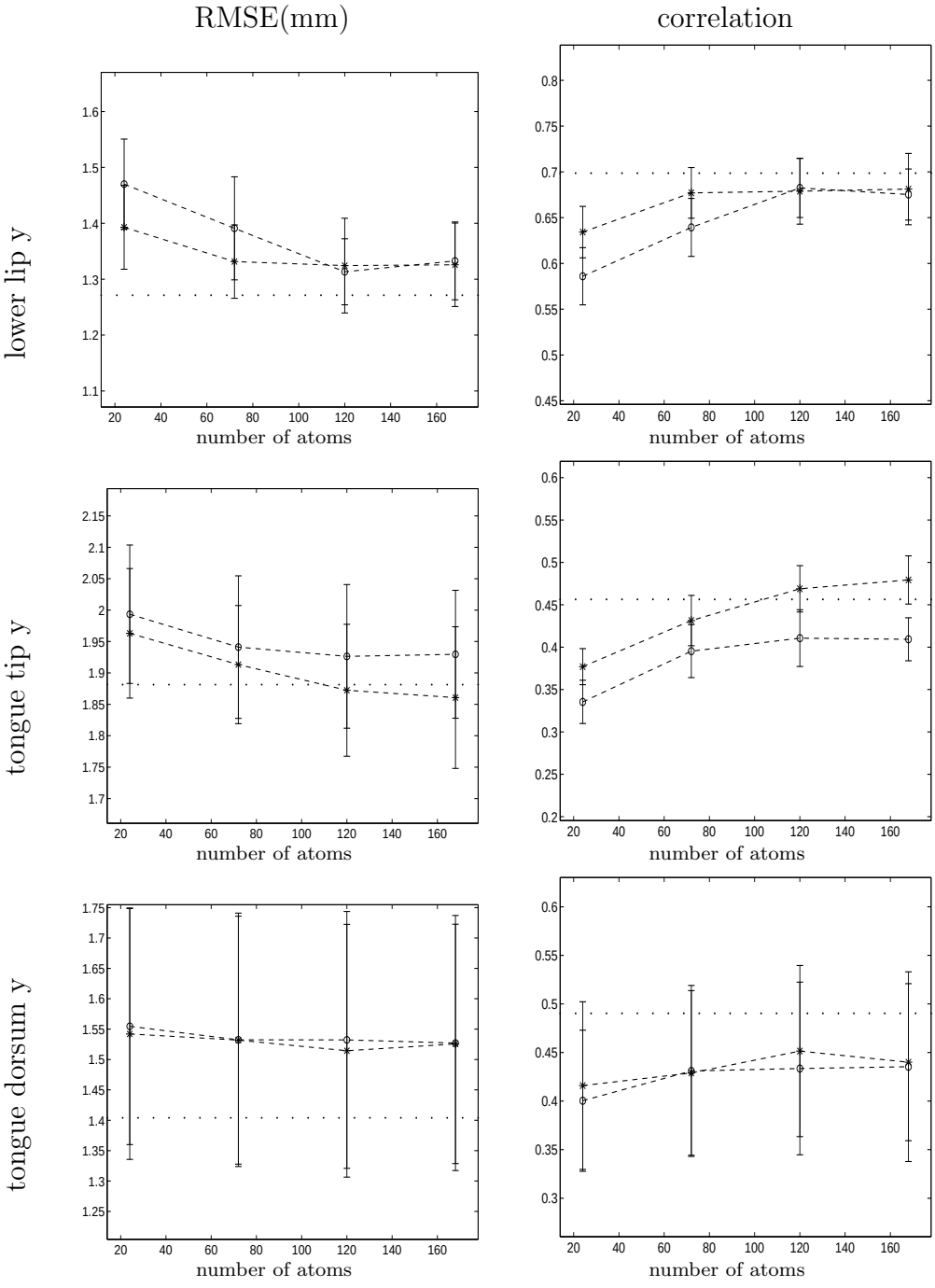


Figure 6-10: Performance in terms of RMSE and correlation using conventional method (noted with ∘) and using relevant time–frequency atoms (noted with ∗) for the critical articulators of the stop consonants. Bilabial (/p, b/), first column of figures; alveolar (/t, d/), second column; and, velar (/k, g/), third column. The ceiling value, best performance among all selected number of frames using conventional method, is depicted as dotted lines in this figure.

	ll_y	tt_y	td_y	total
RMSE improvement (%)	2.3(H_1)	2.3(H_1)	0.5(H_1)	1.7(H_1)
correlation improvement (%)	3.6(H_1)	13.2(H_1)	2.2(H_1)	6.3(H_1)

Table 6-1: Performance improvement in plosives consonants when using relevant maps instead of conventional method for selecting input features. H_1 indicates the null hypthesis is rejected, thus the improvement is statistically significant.

	ul_y	ll_x	ll_y	tt_x	tt_y	td_y
RMSE improvement (%)	0.4(H_0)	1.6(H_1)	-1.2	14.2(H_1)	-3.1	1.3(H_1)
correlation improvement (%)	5.2(H_0)	22.4(H_1)	-2.5	26.7(H_1)	-2.7	4.5(H_1)

Table 6-2: Performance improvement for critical articulatorswhen using relevant maps instead of conventional method for selecting input features. H_1 indicates the null hypthesis is rejected, thus the improvement is statistically significant; by contrast, H_0 means the improvement is not statistically significant.

using 120 relevant inputs compared to 408 inputs required by conventional method, that is a reduction of 70.6 %; b) regarding tt_y, taking 120 relevant TF atoms the results are better than the performance of conventional method, which requires 408 input features.

Acoustic–to–articulatory mapping of critical articulators The improvement percentage obtained by the proposed method in respect to conventional approach, for the articulators ul_y, ll_x, ll_y, tt_x, tt_y and td_y is estimated in a same way as for the preceding section. Diagonal co–variance matrix is used in the models. The results are shown in Figures **6-11** and **6-12**. It can be observed from table **6-2**, which is a compendium of the results shown in Figures **6-11** and **6-12**, that the performance of acoustic–to–articulatory mapping system increases for the articulators ll_x, tt_x and td_y.

6.4. TF relevant features for subject-independent acoustic-to-articulatory mapping of fricatives

Section 6.3 shows the importance of maps of TF relevant features to improving articulatory inversion. Present section discuss the potencial usefulness of same relevant maps, but for subject-independent acoustic-to-articulatory inversion.

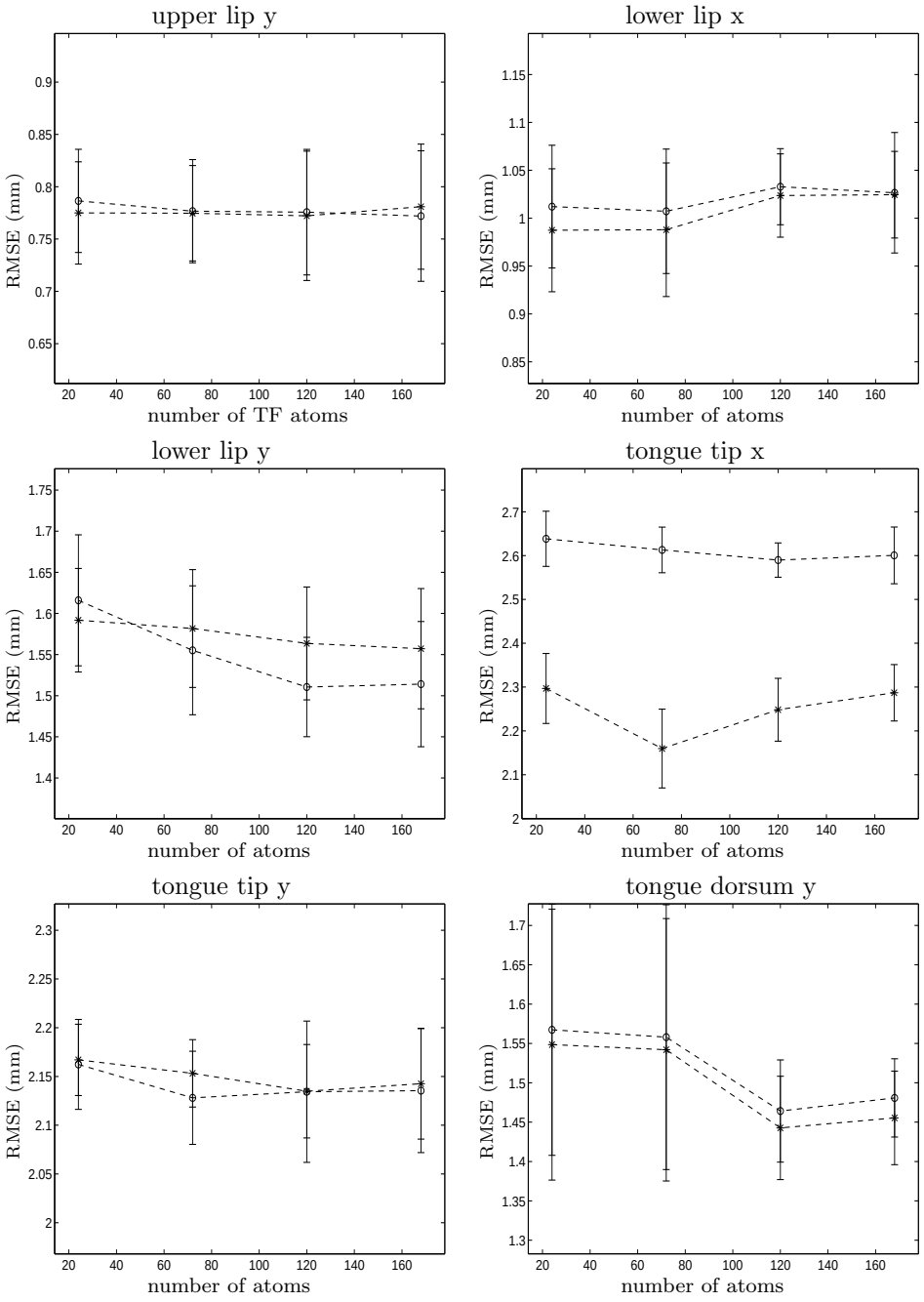


Figure 6-11: Performance in terms of RMSE using conventional method (noted with $\circ$) and using relevant time–frequency atoms (noted with $*$) for the critical articulators.

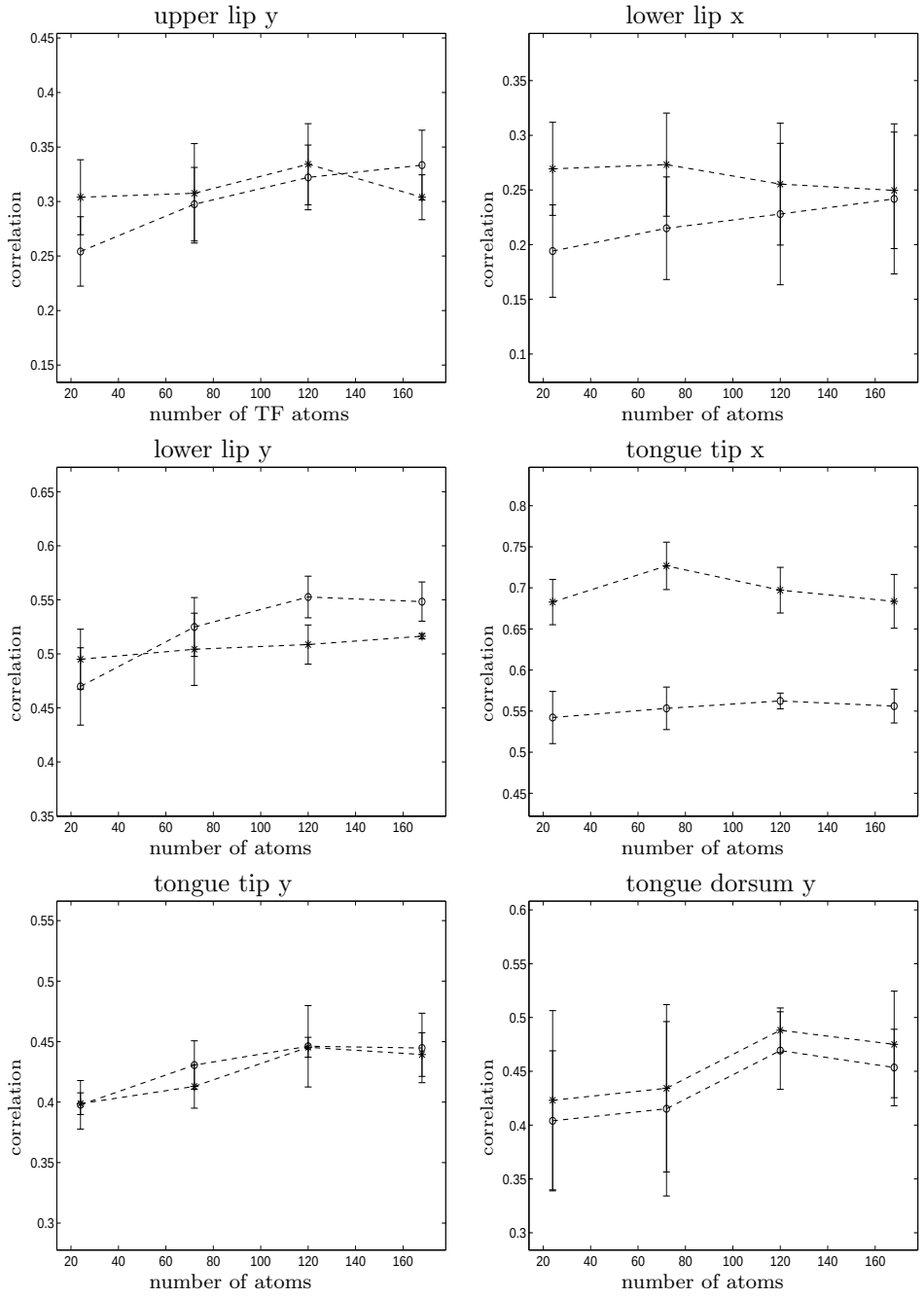


Figure 6-12: Performance in terms of correlation value for critical articulators using conventional method (noted with ∘) and using relevant time–frequency atoms (noted with ∗) for the critical articulators.

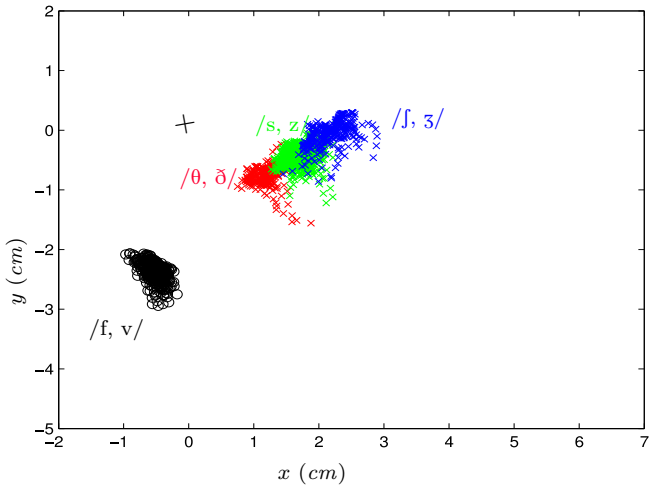


Figure 6-13: Scatter plot of EMA data corresponding to tongue tip (in colors) an lower lip (in black), the critical articulators of fricatives. The plot is obtained by using the phrase 400_{th} to 460_{th} of msak0 speaker.

6.4.1. Speech signal representation

To extract the speech segments corresponding to the fricatives, the labels provided in [43] are used. ll_x and ll_y are critical for phonemes /f, v/; while tt_x and tt_y are critical for fricative phonemes /θ, ð, s, z, ʃ, ʒ/. Thus, two sets of acoustic-articulatory pairs are employed for each speaker. An example of the articulatory data distribution is shown in Figure (**6-13**).

The vocal tract length differences between the female (fsew0) and male (msak0) speakers, of the MOCHA-TIMIT database, are taken into account during the feature estimation procedure. To diminish their influence, vocal tract normalization is performed for both speakers by applying the normalization factors described in [5]. Figure (**6-14**) shows the resulting frequency warping functions that are used for the vocal tract normalization.

In this section, frequency splitting is generated with 24 mel filter banks, as explained in section 2.2.4. To carry out the time plane partition, the acoustic speech signal is parameterized using 20 *ms* frames and $\Delta t = 10$ *ms* steps, so a frame rate of 100 *Hz* is performed. Acoustic information within a time interval ranging from $t - t_a = t - 200$ *ms* to $t + t_b = t + 300$ *ms* is parameterized; thus, a *time–frequency* (TF) plane is obtained. A total of 50 frames taken every 10 *ms* in time are parameterized using the 24 mel filter banks with embedded vocal tract normalization. The TF information is represented by the scalar valued logarithmic energy features $x(t + d, f_k) \in \mathbb{R}$. The matrix of log–energy features is denoted as $\boldsymbol{X}_t \in \mathbb{R}^{N_t \times N_f}$. In this case, $N_c = 4$ channel are analyzed, i. e. tt_x, tt_y, ll_x and ll_y.

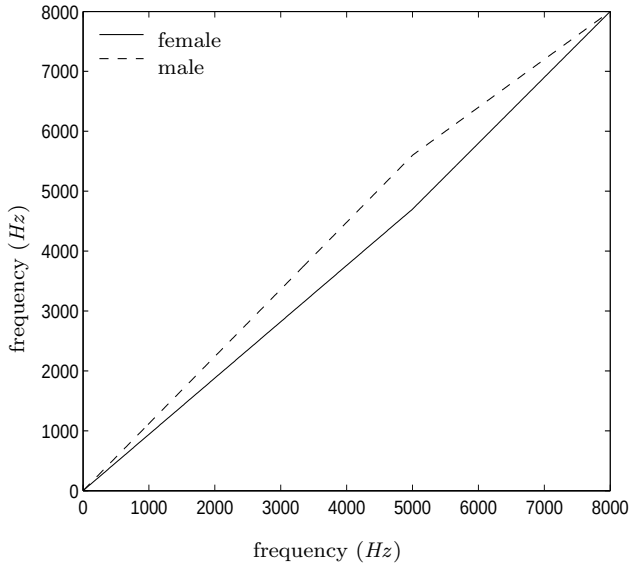


Figure 6-14: Frequency warping functions used for the vocal tract length normalization process of female and male speakers in `MOCHA-TIMIT` database. The slopes of the linear functions beginning in the origin are 0,94 and 1,12 for female and male speakers, respectively.

To estimate the relevant feature set, a statistical measure of association is applied to the TF atoms enclosed within the context window $[t - t_a, t + t_b]$. Particularly, we use the χ^2 information measure $I(x(\cdot), y(\cdot)) \in \mathbb{R}$, as explained in section 3.1.1. Here, the process generates 1200 statistical association values at each time t. A maximum of 2000 pairs $\{\boldsymbol{X}_t,\, y^m(t)\}$ of EMA–acoustic points are taken for the estimation of relevant TF features. The χ^2 information measure coefficient is carried out between each variable $x(t + d, f_k)$ and articulatory trajectories of the four corresponding EMA channels. The resulting points are used to build the relevant TF feature set.

6.4.2. Inversion of fricatives using relevant TF features

The relevant time–frequency atoms without VTLN in case of fricatives /θ, ð, s, z, ʃ, ʒ/ for male a female speaker are shown in figure **6-15**. Even though a frame step of 18 ms is utilized in [77], instead of of using 10 ms time–shift; we selected 10 ms shift rate because it is more widely used. The number of inputs is varied ranging from $p = 24$ to $p = 120$ ($p = 24, 72$, and 120); that is, 1, 3, and 5 frames around current time of analysis are taken into account. The input vector is transformed using Principal Component Analysis, where $n_p = 24, 35, 35$ components are taken, respectively. In the case of relevant maps,

the $p = 24, 72$ and 120 most relevant atoms are used. Then, the $n_p = 24, 35, 35$ principal components are extracted to form the input vector for the model in (3-27). In all cases 32 mixtures are used. The model parameters are found by using the expectation maximization (EM) algorithm. It must be quoted that the articulatory estimations are not low–pass filtered in this work. To measure the accuracy of the mapping a 5–fold cross–validation testing is carried out. The 460 sentences are divided into 5 partitions consisting of 92 sentences, and then one of the partitions is reserved for testing by turns, while the other 4 partitions are used for training. The performance is measured by using the root mean square error and the Pearson's correlation coefficient.

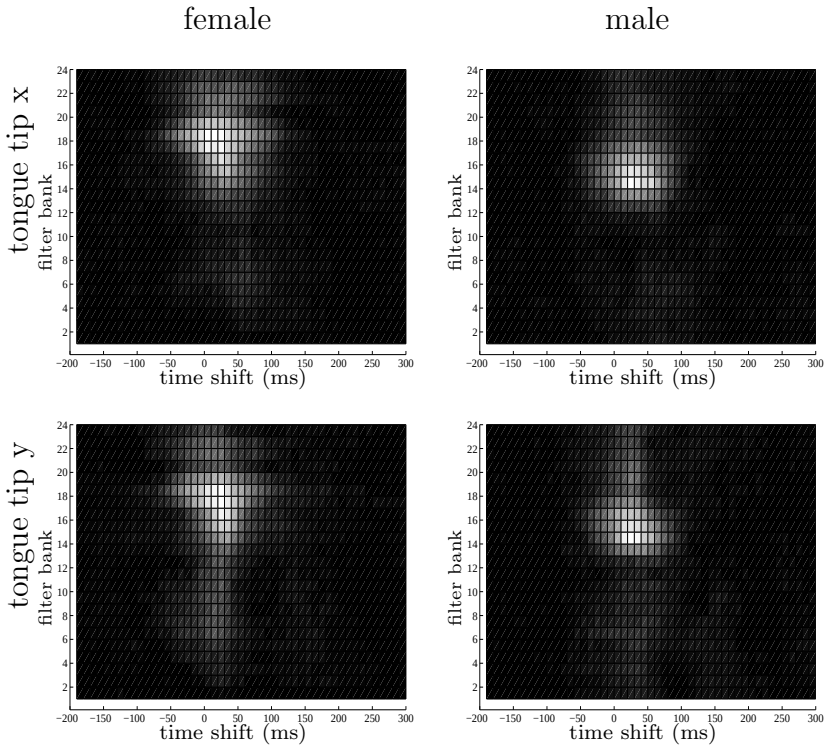


Figure 6-15: Relevant time–frequency atoms without VTLN for the critical articulators of the fricative phonemes /θ, ð, s, z, ʃ, ʒ/. (tt_x and tt_y).

For each of the 5 partitions (consisting of 92 sentences) the phones corresponding to fricative phonemes are extracted and used to evaluate the relevant features. One of the sets is reserved for testing by turns, while the other 4 sets are used for training. For the sake of avoiding any possible problem caused by reduced number of samples available for training and testing processes, we choose diagonal co-variance matrix. The results, in terms of average RMSE and average correlation between both speakers, are shown in Figure (**6-16**). It can be observed that the performance of acoustic–to–articulatory mapping system increases for

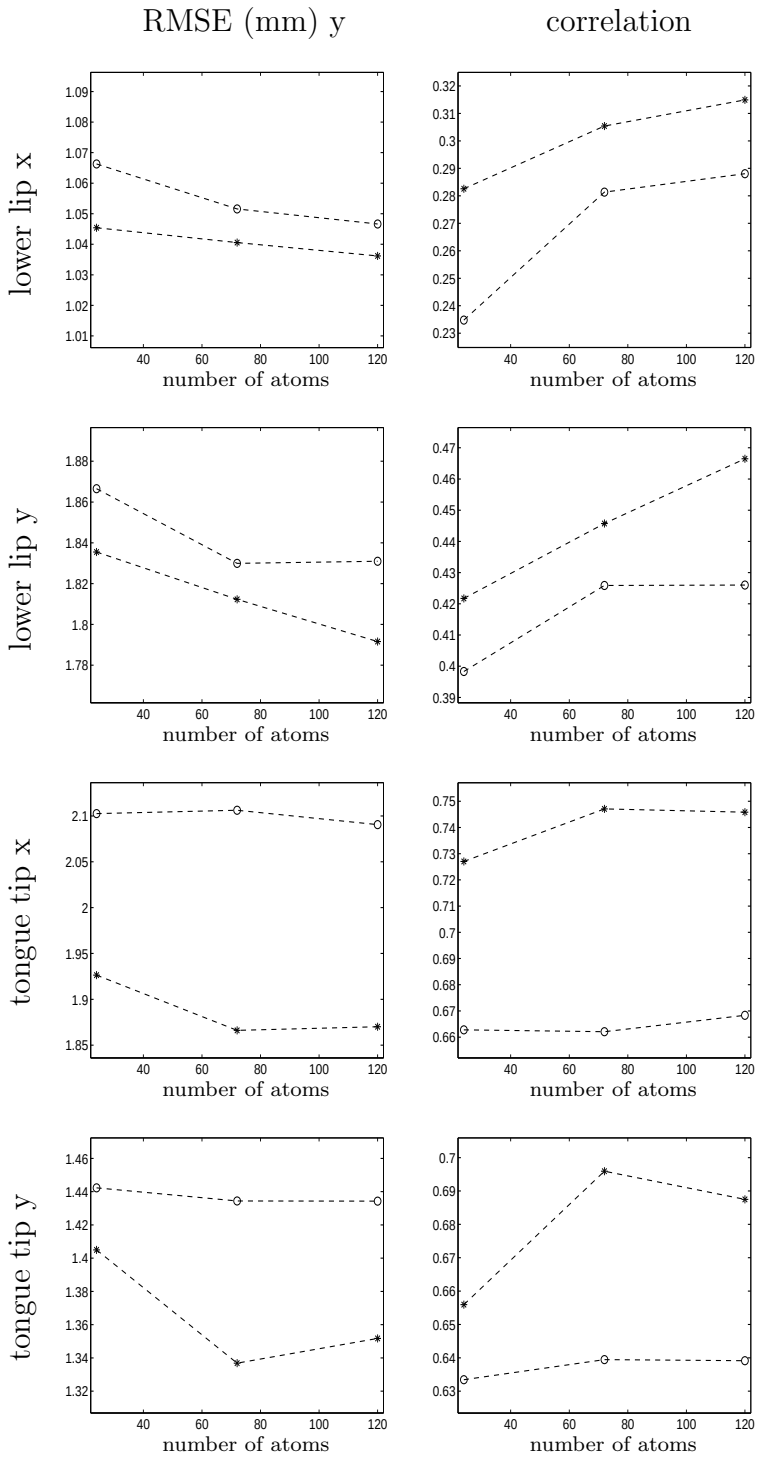


Figure 6-16: Performance in terms of RMSE and correlation using conventional method (noted with ∘) and using relevant time–frequency atoms (noted with ∗) for the critical articulators of fricative consonants.

the articulators involved in the production of fricatives.

6.4.3. Subject-independent inversion of fricatives using relevant TF features and VTLN

The relevant features are estimated for the female and male speakers of the MOCHA database (note that there are only two subjects in the corpus). As seen in Figure (**6-17**) showing the relevant TF atoms corresponding to fricatives /θ, ð, s, z, ʃ, ʒ/ and /f, v/, the relevant zones for the channel ll_x are very diffuse. This observation agrees with the distribution of lower lip shown in Figure (**6-13**), where one can see that the major part of the variance is along y axis.

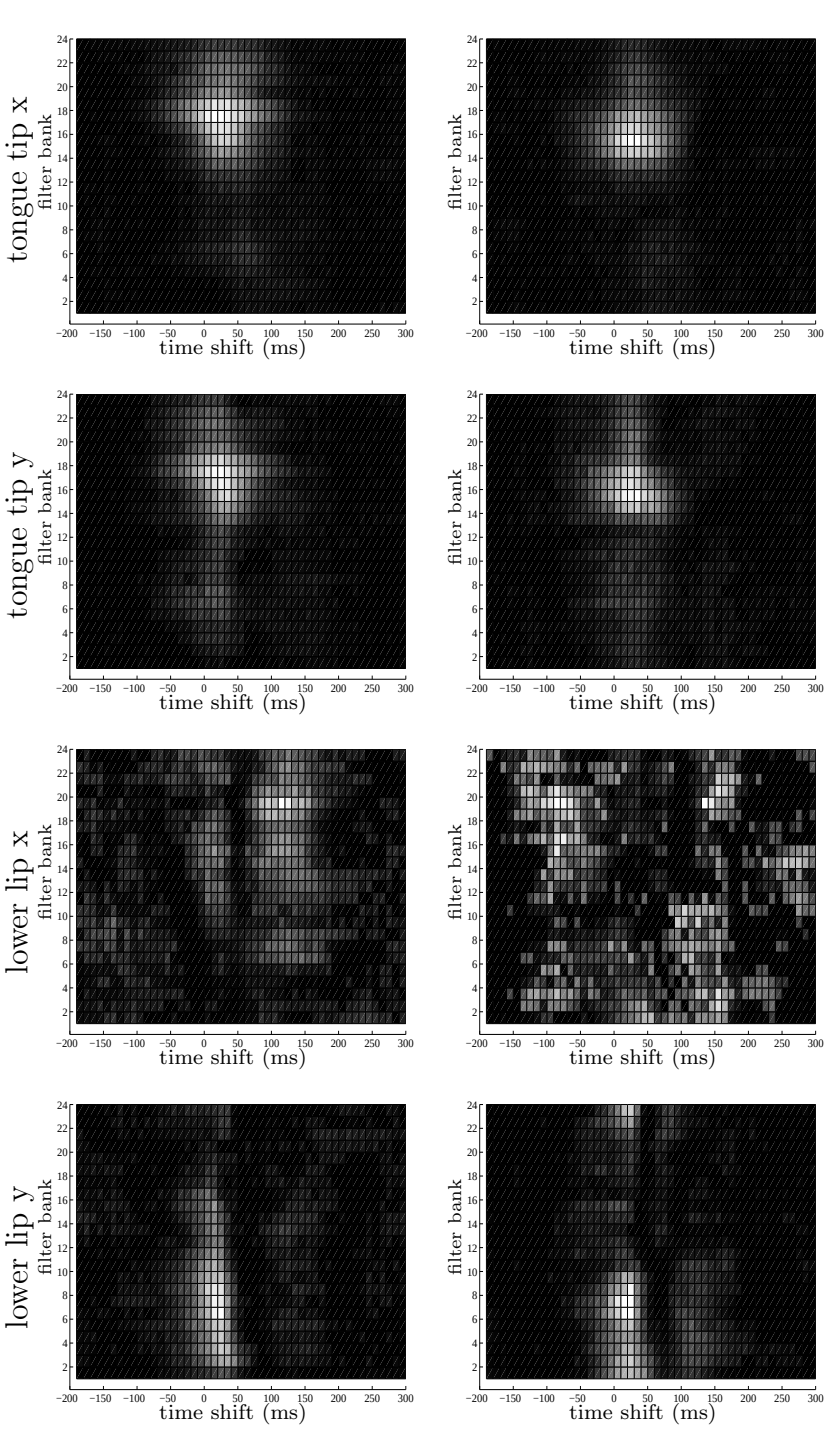


Figure 6-17: Relevant time–frequency atoms for the critical articulators of the fricative phonemes /θ, ð, s, z, ʃ, ʒ/ (tt_x and tt_y); and /f, v/ (ll_x and ll_y). The maps are obtained after applying VTLN process by using the warping functions shown in Figure (**6-14**).

Within the experimental framework, we consider three approaches to be compared: a) the proposed inversion method (noted as IA-1), which makes use of relevant TF features and vocal tract length normalization (VTLN) procedure; b) the IA-2 approach corresponding to the conventional subject-dependent method used in previous works [90, 110], and c) the IA-3 approach that is similar to the IA-2 except that the training data is obtained from one subject while the other subject's data is used for testing.

In case of IA-2 and IA-3 approaches, the number of inputs ranges from $p = 24$ until $p = 120$ ($p = 24, 72$, and 120); that is, 1, 3, and 5 frames around current time of analysis were taken into account. The input vector was projected using Principal Component Analysis, where $n_p = 24, 35, 35$ components were taken, respectively. When employing the IA-1 approach, the $p = 24, 72$, and 120 most relevant atoms were used. Then, the $n_p = 24, 35, 35$ principal components were extracted to form the input vector for the model given in Section 3.1.2. In all cases, 32 mixtures were used. The model parameters were found by using the expectation maximization algorithm. To measure the accuracy of the mapping, a 5–fold cross–validation testing was carried out. The 460 sentences were divided into 5 partitions consisting of 92 sentences, and then one of the partitions was reserved for testing by turns, while the other 4 partitions were used for training. The articulatory estimations were not low–pass filtered in this work.

Although for the same articulator, the raw EMA data may vary between subjects. However, the shape of articulatory trajectories is expected to be similar for each one of the phoneme [32]. Therefore, correlation value is used to measure the inversion quality. The performance for each selected number of input features is assessed by using the Pearson's correlation coefficient, $\hat{\rho}$, which consists of the average correlation along the number inputs; that is, the average among the obtained values when using $p = 24, 72$, and 120 features. As seen in Figure (**6-18**) showing the $\hat{\rho}$ values for the msak0 and fsew0 speakers, the proposed method offers a better performance in respect to IA-3, and it is comparable to inversion scheme IA-2.

6.5. Discussions

On estimating relevant features From the estimated relevance maps, see Figures **6-1**, **6-2**, **6-3**, it can be observed that the zones of maximal association values are located after the current time of analysis, i.e., following the temporal position of the articulatory information, for the majority of articulators analyzed in present work. The relationship between the

position of maximal relevance zones and the articulator configuration is fairly complex, and its explanation is out of the scope of this paper. The zones of maximal information tends to be located on lower ranges of frequency for male speaker in respect female speaker, but preserving the similarities in shape, particularly, when modelling tongue tip x and tongue tip y, see Figures **6-2**, **6-3**. Observing Figures **6-1**, **6-2**, **6-3** additional similarities can be appreciated between the relevance maps of the female speaker and the male speaker; though, not for all articulators.

The burst spectrum of stops can be used as a cue to place [50]. In previous works it has been found that labials tend to have diffusely falling or flat spectra, alveolars have diffusely rising patterns, and velars exhibit higher concentration of energy in the intermediate frequency and relatively low–frequency regions. We found some similarities between previous works on acoustics phonetics [50] and achieved time–frequency relevant maps, namely: a) in case of bilabials, for female speaker relevant components are flatly distributed along a considerable part of the spectrum and for the male speaker this pattern is falling; b) for alveolar stop phones, the relevant atoms are almost concentrated on high–frequency components, for female as well as male speakers; and c) for velar stops, the relevant components of fsew0 are compactly distributed around relatively low–frequency components.

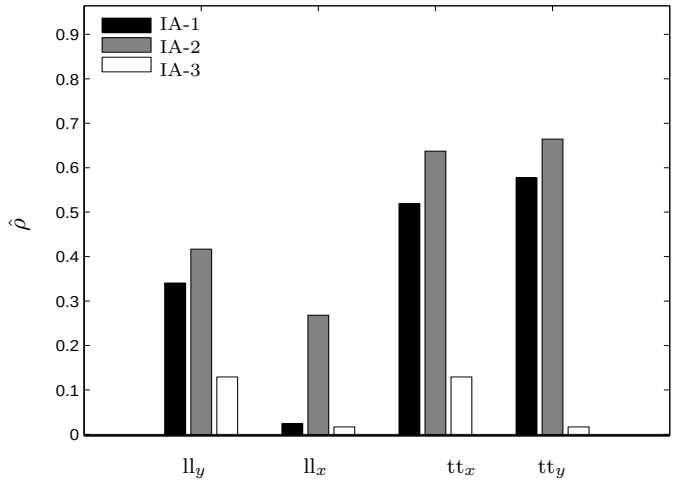
Assesment of Kendall measure for measuring the statistical dependence between articulators and speech components In case of analysis of the relationship between articulators and the TF information, the Kendall τ coefficient has shown some useful properties. Mainly, the Kendall coefficient can detect most of the relations, as shown in (6.2). As well as being useful for estimating the articulatory information distribution, the simplicity offered by the Kendall coefficient (as seen in Eq. (3-3)) makes it easy to implement. Even though Kendall coefficient and partial correlation Kendall coefficient are simpler measures of statistical association in respect to other measures of statistical association, they are able to detect most of the relationships between articulatory positions and the time-frequency acoustic information, as observed in Figures **6-6** and **6-7**. Theses results act as a complement to the one shown in **6-4**.

Relevant acoustic-to-articulatory maps using Gaussian mixture regression It is observed in [110] that the performance improves as the number of input features increases, where the best performance is achieved when using 256 input features. In present work, if assuming diagonal matrix for the model covariance and conventional method in case of stops,

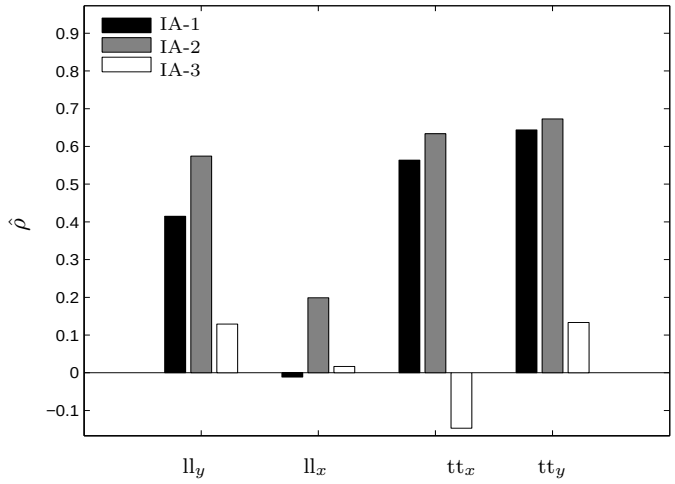
the best performance is obtained when using 408 features for modelling ll_y and tt_y. We name this value p^*. By contrast, using the proposed relevant features, the performance is close to p^* in case of ll_y, while for tt_y, the performance is overpassed; but with the benefit that in both cases (see Figure **6-10**), instead of using 408 features only 120 relevant TF atoms are necessary. Results shown in Table **6-2** also support the usefulness of proposed method; where it is observed the improvement in terms of correlation and RMSE of some critical articulators analyzed, when using relevant features in respect to conventional method.

However, it can be seen that there is not advantage on using relevant features in respect to using conventional method, in case of whole utterance based analysis. This phenomenon could be caused by the fact that relevant maps are estimated along whole utterance. That is, it includes phones, for which a given articulator is critical, as well as phones where the same articulator is also not critical.

Subject-independent inversion of fricatives using relevant TF features and VTLN Regarding subject-independent inversion of fricatives, the Figure (**6-18**) is obtained by using the frequency warping functions of Figure (**6-14**) with slopes 0,94 and 1,12 for female and male speakers, respectively. This parameter turns to be key for the performance of the system, as shown in Figure (**6-19**). In case of male speaker, it can be observed that the best performance occurs approximately at $\alpha = 1{,}12$, whereas regarding a female speaker, the best VTLN parameter is $\alpha = 0{,}9$ instead of $\alpha = 0{,}94$. Using $\alpha = 0{,}9$ might cause the most relevant features for fsew0 speaker be moved to lower frequency regions.

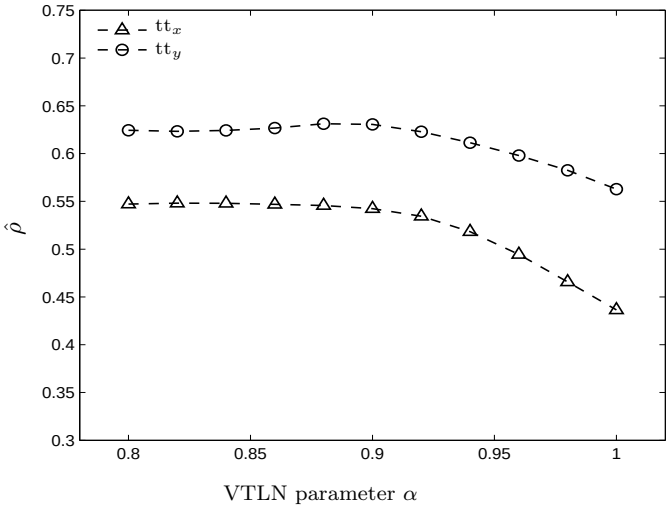


(a) Performance $\hat{\rho}$ for the speaker fsew0

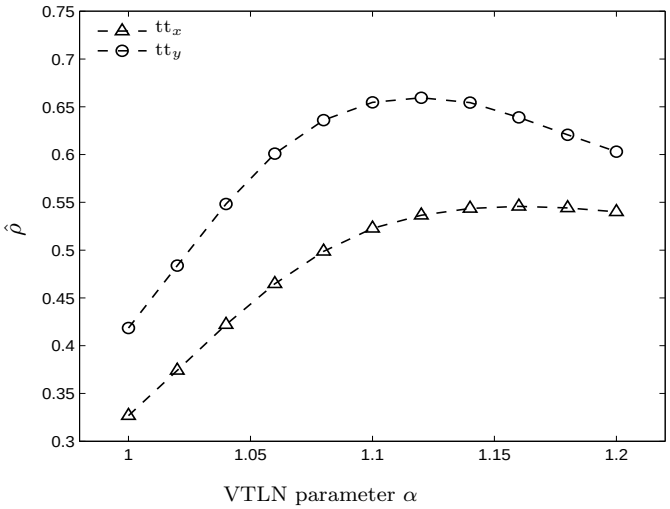


(b) Performance $\hat{\rho}$ for the speaker msak0

Figure 6-18: Correlation average performance $\hat{\rho}$ for speakers fsew0 and msak0 by using the inversion appraches IA-1, IA-2 and IA-3. IA-1 is the proposed subject-independent inversion scheme; IA-2 is the subject-dependent inversion method commonly used in recent works.



(a) $\hat{\rho}$ versus VTLN parameter α for fsew0 speaker



(b) $\hat{\rho}$ versus VTLN parameter α for msak0 speaker

Figure 6-19: Correlation average performance $\hat{\rho}$ for tongue tip (tt_x and tt_y) of speakers fsew0 and msak0.

7 Conclusions

Conclusions

The use of measures of statistical association for the feature selection process leads to the improvement in performance of acoustic-to-articulatory mapping systems, particularly those based on Gaussian mixture models. Relevant maps are estimated over the whole speech signal as well as on specific phones, for which a given articulator is known to be critical. It is important to remark that in case of whole utterance based analysis, there is not noticeable advantage on using relevant features in respect to using inputs conventionally selected. However, the relevant maps are useful for the position inference of critical articulators, as exposed in section 6.1. In addition, same relevant maps based method is useful for the subject-independent acoustic-to-articulatory mapping, particularly in case of fricative phonemes. It is important to point out that the vocal tract normalization process on the speech signal plays a crucial role, as shown in the same section.

However, as shown in present work, not only statistically relevant features are useful to improve the performance. Among those potential acoustic parameters, the use of formants were analyzed in present work. Results display that the statistical relation between formants and the form of the vocal tract is higher than for the case of MFCC. Then, it is found that this higher statistical relation leads to a more precise estimation of articulator position; especially the tongue tip, blade and body.

In addition, this work shows that using the adequate input features makes possible the inference of critical articulators movement in a subject-independent way. In our case, we utilized time-frequency relevant features after vocal tract normalization process. With respect to critical articulators movement inference, as a result, a set of invariant acoustic features for fricatives is obtained.

On the other hand, because of the nature of input selection strategy, the relevant maps provide a deeper understanding into the relationship between the articulatory and acoustical

phenomena. The relevant maps give the location in time and frequency of those features being mostly statistically related to articulators.

Regarding analysis-by-synthesis methods, a optimization procedure was used to seek a lifter such that it could be used used to replace formants in analysis-by-synthesis acoustic-to-articulatory inversion tasks. The derived liftered cepstral distance is able to suppresses sufficiently glottal variability so that a minimal distance in the cepstral domain implies a minimal distance in the formant domain may be achieved by using proposed method. The lifter is shown to outperform other lifters previously presented in the literature.

Future work

Multispeaker inversion mapping Perhaps the most interesting aspects of subject-independent inversion mapping is to handle speakers for whom there is not articulatory training data. For the end user, collecting EMA data of their speech for further training purposes is impractical. It would be captivating to research and develop a generic inversion mapping strategy with the capability to infer articulatory parameters of any individual without the use of further EMA data. However, several complications should be solved first. Although there are many solutions to a particular problem, in this thesis we propose the following steps.

- The proposed method, corresponding to using time-frequency relevant maps with vocal tract length normalization, should be extended to other manner of articulation categories (stops, nasals, vowels and semivowels). Then; the resulting models would be mixed with a segmentation system of proven performance, thus helping to obtain a system that works on the entire speech signal.
- Obtain some kind of standard representation of the vocal tract, such that it is capable of representing the salient properties of the vocal tract of the speakers, but with minor adjustments. A candidate to meet this requirement is the representation of Maeda, but first it is required to use a model that relates data from EMA systems with Maeda's model. This problem has been addressed recently by Toutios et. al. in [113].
- At this point, the trajectories of critical articulators would have been obtained. Using the speech signal to estimate all articulator movement at all times and in a subject-independent way may not be an attainable goal. However, the existence of inter-articulator correlation phenomenon is well-known; and it could be used, with prior

estimation of critical articulators trajectories, to infer the trajectories of non-critical articulators.

Speech therapy The degree of hypernasality is related to the area of the velopharyngeal gap; and, this area is directly related to the velum position during the generation of speech sounds [56]. An effective acoustic-to-articulatory mapping system would be helpful for estimating the velopharyngeal gap area. Another problem consist on the shift in localization of articulation that occurs in cleft palate children after reconstructive surgical treatment; particularly in fricative sounds [71, 22]. This problem can be treated from the perspective of acoustic-to-articulatory inversion.

Use of more articulatory data The use of EMA data have been becoming more popular. The authors suggest applying proposed relevant TF features to an articulatory database with a greater number of speakers in order to go beyond in the understanding of the relationship between the vocal tract shape and the acoustic speech signal. Moreover, it is proposed to create a database that helps to the study of Spanish language.

Most informative TF atoms Another aspect to take into account for future work is on measurements on the relative advantages of relational spectral cues over single spectral cues, as follows: the most informative TF atom will be fixed; then, conditional relevant maps are computed in order to estimate the additional information provided by a new feature located in the vicinity of the given feature. The process will generate a second most informative feature conditioned to the fact we had previously selected the most informative coordinate. This process will be repeated to find optimal three–points, perhaps four–points, and so on.

References

[1] *X-ray microbeam speech production database user's handbook version 1.0*.

[2] Ahmed Abdelatty, Jan Van der Spiegel, and Paul Mueller. Acoustic-phonetic features for the automatic classification of fricatives. *J. Acoustical Society of America*, 109(5), May 2001.

[3] Paul S. Addison. *The illustrated wavelet transform handbook*. Institute of Physics Publishing, 2002.

[4] Ali N. Akansu and Richard A. Haddad. *Multiresolution Signal Decomposition : Transforms, Subbands, and Wavelets*. Academic Press, second edition, 2001.

[5] Ziad Al Bawab. *An Analysis-by-Synthesis Approach to Vocal Tract Modeling for Robust Speech Recognition*. PhD thesis, Carnegie Mellon University, 2009.

[6] Samer Al-Moubayed and G. Ananthakrishnan. Acoustic–to–articulatory inversion based on local regression. In *InterSpeech–2010*, 2010.

[7] G. Ananthakrishnan and Olov Engwall. Mapping between acoustic and articulatory gestures. *Speech Communication*, 53(4):567–589, 2011.

[8] G. Ananthakrishnan, Daniel Neiberg, and Olov Engwall. In search of non-uniqueness in the acoustic-to-articulatory mapping. In *InterSpeech*, 2009.

[9] Peter F. Assmann. The role of formant transitions in the perception of concurrent vowels. *J. of Acoustical Society of America*, 97:1, 1995.

[10] Pierre Badin, Yuliya Tarabalka, Frederic Elisei, and Gerard Bailly. Can you read tongue movements? : Evaluation of the contribution of tongue display to speech understanding. *Speech Communication*, 52(6):493–503, 2010.

[11] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[12] Jun Cai, Yves Laprie, Julie Busset, and Fabrice Hirsch. Articulatory modeling based

on semi-polar coordinates and guided PCA technique. In *InterSpeech*, pages 56–59, 2009.

[13] Donald G. Childers, David P. Skinner, and Robert C. Kemerait. The cepstrum : A guide to processing. *Proceedings of the IEEE*, 65(10), October 1977.

[14] Ingrid Daubechies. *Ten Lectures on Wavelets*. SIAM: Society for Industrial and Applied Mathematics, 1992.

[15] Genaro Daza, Luis Sanchez, Alexander Sepulveda, and German Castellanos. *Encyclopedia of Healthcare and Information Systems*, chapter Acoustic Feature Analysis for Hypernasality Detection in Children. 2008.

[16] J. R. Deller, J. H. Hansen, and J. G. Proakis. *Discrete-time processing of speech signals*. John Wiley & Sons, 1993.

[17] Lee Deng and Douglas O'Shaughnessy. *Speech processing: a dynamic and optimization-oriented approach*. Marcel Dekker, Inc., 2003.

[18] Jean Dickinson and Subhabrata Chakraborti. *Nonparametric Statistical Inference*. Marcel Dekker, Inc., 4 edition, 2003.

[19] Richard Duda, Peter Hart, and David Stork. *Pattern Classification*. John Wiley & Sons, second edition, 2001.

[20] Sorin Dusan and Li Deng. Recovering vocal tract shapes from MFCC parameters. In *ICSLP*, 1998.

[21] O. Engwall. Vocal tract modelling in 3D. *TMH-QPSR*, 40(1-2):31–38, 1999.

[22] M. L. Falk and G. A. Kopp. Tongue position and hypernasality in cleft palate speech. *The Cleft Palate Journal*, 5(3):228–237, 1968.

[23] O. Farooq and S. Datta. Mel filter-like admissible wavelet packet structure for speech recognition. *IEEE Signal Processing Letters*, 8:196–198, 2001.

[24] Daniel Felps and et. al. Relying on critical articulators to estimate vocal tract spectra in an articulatory-acoustic database. In *InterSpeech*, pages 1990–1993, 2010.

[25] J. Frankel and S. King. Speech recognition using linear dynamic models. *IEEE Transactions on Audio, Speech and Language Processing*, 15(1):246–256, 2007.

[26] Joe Frankel, Mirjam Wester, and Simon King. Articulatory feature recognition using dynamic bayesian networks. *Computer Speech and Language*, Octuber 2007.

[27] Sadaoki Furui. *Digital Speech Processing Synthesis and Recognition*. Marcel Dekker Inc., 1989.

[28] and Olov Engwall G. Ananthakrishnan and Daniel Neiberg. Exploring the predictability of non-unique acoustic-to-articulatory mappings. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(10):2672–2682, December 2012.

[29] O. Ghitza and J. Goldstein. Scalar LPC quantization based on formant JND's. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):697–708, 1986.

[30] Prasanta Kumar Ghosh, Louis M. Goldstein, and Shrikanth Narayanan. Processing speech signal using auditory-like filterbank provides least uncertainty about articulatory gestures. *J. Acoust. Soc. Am.*, 129(6), June 2011.

[31] Prasanta Kumar Ghosh and Shrikanth Narayanan. A generalized smoothness criterion for acoustic-to-articulatory inversion. *Journal of Acoustical Society of America*, 128(4):2162–2172, 2010.

[32] Prasanta Kumar Ghosh and Shrikanth Narayanan. A subject-independent acoustic-to-articulatory inversion. In *ICASSP*, 2011.

[33] A. Gray and J. Markel. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(5):380–391, 1976.

[34] Mirko Grimaldi and et. al. New technologies for simultaneous acquisition of speech articulatory data: 3d articulograph, ultrasound and electroglottograph. In *LangTech*.

[35] Rodrigo Capobianco Guido and et. al. A new technique to construct a wavelet transform matching a specified signal with applications to digital, real time, spike, and overlap pattern recognition. *Digital Signal Processing*, 16:24–44, 2006.

[36] Isabelle Guyon and André Elisseeffe. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[37] Mark Hasegawa-Johnson. Time–frequency distribution of partial phonetic information measured using mutual information. In *InterSpeech–2000*, pages 133–136, 2000.

[38] Sadao Hiroya and Masaaki Honda. Estimation of articulatory movements from speech acoustics using an hmm-based speech production model. *IEEE Transactions on Audio, Speech, and Language Processing*, 12(2):175–185, March 2004.

[39] Sadao Hiroya and Takemi Mochida. Multi-speaker articulatory trajectory formation based on speaker-independent articulatory HMMs. *Speech Communication*, 48:1677–

1690, 2006.

[40] John Hogden, Anders Lofqvist, Vince Gracco, Igor Zlokarnik, Philip Rubin, and Elliot Saltzman. Accurate recovery of articulator positions from acoustics: new conclusions based on human data. *Journal of Acoustical Society of America*, 100(3):1819–1834, September 1996.

[41] X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall PTR, 2001.

[42] F. Itakura and T. Umezaki. Distance measure for speech recognition based on the smoothed group delay spectrum. In *ICASSP '87*, Dallas TX, April 1987.

[43] Philip Jackson and Veena Singampalli. Statistical identification of articulation constraints in the production of speech. *Speech Communication*, 51(8), 2009.

[44] Julie Fontecave Jallon and Frédéric Berthommier. Asemi-automaticmethod for extractingvocaltractmovements from x-rayfilms. *Speech Communication*, 51(2):97–115, February 2009.

[45] B. H. Juang, L. Rabiner, and J. Wilpon. On the use of band-pass liftering in speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(7), July 1987.

[46] Daniel Jurafsky and James Martin. *Speech and Language Processing*. Prentice Hall, 2000.

[47] Erwin K. *Introductory Functional Analysis with Applications*. John Wiley & Sons, 1989.

[48] H. Kadri. Learning vocal tract variables with multi-task kernels. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[49] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

[50] Ray Kent and Charles Read. *Acoustic Analysis of Speech*. Delmar–Thomson Learning, second edition, 2002.

[51] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester. Speech production knowledge in automatic speech recognition. *Journal of Acoustical Society of America*, 121(2):723–742, February 2007.

[52] Hedvig Kjellström and Olov Engwall. Audiovisual-to-articulatory inversion. *Speech Communication*, 51(3):195–209, 2009.

[53] T. G. Kolda, Robert M. L., and Virginia Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *Society for Industrial and Applied Mathematics Review*, 45(3):385–482, 2003.

[54] A. N. Kolomogorov and S. V. Fomin. *Introductory Real Analysis*. Dover Publications, Inc., 1970.

[55] D. Kouamé and et. al. Ultrasound imaging: signal acquisition, new advanced processing for biomedical and industrial application. In *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*.

[56] Ann Kummer. *Cleft Palate & Craniofacial Anomalies*. Delmar–Thomson Learning, 2001.

[57] J. Lagarias, J. Reeds, M. Wright, and P. Wright. Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9(1):112–147, 1998.

[58] Y. Laprie and M.-O. Berger. Cooperation of regularization and speech heuristics to control automatic formant tracking. *Speech Communication*, 19:255–269, 1996.

[59] Yves Laprie. Snorri, a software for speech sciences. In *ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education MATISSE*, pages 89–92, 1999.

[60] Yves Laprie and et. al. Acoustic-to-articulatoy inversion: Methods and acquisition of articulatory data. Technical report, ASPI consortium: Audiovisual to Articulatory Speech Inversion, 2006.

[61] Yves Laprie and et. al. Audio-visual to articulatory speech inversion: final report on evaluation results. Technical report, ASPI consortium: Audiovisual to Articulatory Speech Inversion, 2009.

[62] and Alex Acero Li Deng and Issam Bazzi. Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint. *IEEE Transactions On Audio, Speech, And Language Processing*, 14(2), March 2006.

[63] R. Pruvenok J. Huang S. Momen Y. Chen Li Deng, X. Cui and A. Alwan. A database of vocal tract resonance trajectories for research in speech processing. In *Int. Conf. on*

Acoustics, Speech, and Signal Processing, May 2006.

[64] Bjorn Lindblom and John Sundberg. Acoustic consequences of lip, tongue, jaw and larynx movement. *J. of Acoustical Society of America*, 50, 1971.

[65] Zhen-Hua Ling, Korin Richmond, and Junichi Yamagishi. An analysis of HMM-based prediction of articulatory movements. *Speech Communication*, 52:834–846, 2010.

[66] Ian R. A. MacKay. *Phonetics: the science of speech production*. Colege Hill, second edition, 1987.

[67] Shinji Maeda. Un modele articulatoire de la langue avec des composantes lineaires. In *10émes Journées d'études sur la parole*, pages 152–162, 1979.

[68] Shinji Maeda. A digital simulation method of the vocal-tract system. *Speech Communication*, 1(3-4):199–229, 1982.

[69] Shinji Maeda. Compensatory articulation in speech: analysis of x-ray data with an articulatory model. In *First European Conference on Speech Communication and Technology, EUROSPEECH '89*, pages 2441–2444, 1989.

[70] Shinji Maeda. *Speech Production and Speech Modelling*, chapter Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using articulatory model, pages 131–149. Kluwer Academic Publishers, 1990.

[71] Andreas Maier and et. al. PEAKS–a system for the automatic evaluation of voice and speech disorders. *Speech Communication*, 51:425–437, 2009.

[72] Pradipta Maji. f-information measures for efficient selection of discriminative genes from microarray data. *IEEE Transactions on Biomedical Engineering*, 56(4):1063–1069, April 2009.

[73] Stéphane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

[74] D. Mansour and B. H. Juang. A family of distortion measures based upon projection operation for robust speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(11):1659–1671, 1989.

[75] P. Meyer, J. Schroeter, and M. M. Sondhi. Desing and evaluation of optimal cepstral lifters for accessing articulatory codebooks. *IEEE Transactions on Signal Processing*, 39(7):1493–1502, 1991.

[76] Slim Ouni and Yves Laprie. Modeling the articulatory space using a hypercube code-

book for acoustic-to-articulatory inversion. *Journal of Acoustical Society of America*, 118(1):444–460, 2005.

[77] Yucel Ozbek, Mark Hasegawa-Johnson, and Mubeccel Demirekler. Estimation of articulatory trajectories based on gaussian mixture model (gmm) with audio-visual information fusion and dynamic kalman smoothing. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(11), July 2011.

[78] S. Panchapagesan and A. Alwan. Vocal tract inversion by cepstral analysis-by-synthesis using chain matrices. In *Interspeech '08*, Brisbane, Australia, September 2008.

[79] Sankaran Panchapagesan and Abeer Alwan. A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the maeda articulatory model. *J. Acoust. Soc. Am.*, 129(4):2144–2162, 2011.

[80] George Papcun and et. al. Inferring articulation and recognizing gestures from acoustics with a neural network trained on x–ray microbeam data. *Journal of Acoustical Society of America*, 2, August 1992.

[81] B. Potard and Y. Laprie. Compact representations of the articulatory-to-acoustic mapping. In *Interspeech '07*, pages 2481–2484, Antwerp, Belgium, August 2007.

[82] B. Potard, Y. Laprie, and S. Ouni. Incorporation of phonetic constraints in acoustic-to-articulatory inversion. *Journal of Acoustical Society of America*, 123(4):2310–2323, 2008.

[83] Tarun Pruthi and Carol Espy-Wilson. Acoustic parameters for automatic detection of nasal manner. *Speech Communication*, 43:225–239, 2004.

[84] C. Qin and M. Á. Carreira-Perpiñán. An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping. In *InterSpeech*, pages 74–77, 2007.

[85] Chao Qin and Miguel A. Carreira-Perpi nan. Reconstructing the full tongue contour from EMA-X-ray microbeam. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4190–4193, 2010.

[86] L. Rabiner and R. Schafer. Introduction to digital speech processing. *Foundations and Trends in Signal Processing*, 1(1-2):1–194, 2007.

[87] S. Rao. *Engineering Optimization: Theory and Practice*. Jhon Wiley, 1996.

[88] Korin Richmond. A trajectory mixture density network for the acoustic-articulatory inversion mapping. In *Interspeech*.

[89] Korin Richmond. *Estimating Articulatory Parameters from the Acoustic Speech Signal*. PhD thesis, The Centre for Speech Technology Research, Edinburgh University, 2001.

[90] Korin Richmond, Simon King, and Paul Taylor. Modelling the uncertainty in recovering articulation from acoustics. *Computer, Speech & Language*, 17:153–172, 2003.

[91] Thomas Ryan. *Moder Regression Methods*. John Wiley & Sons, 1997.

[92] F. Schoen. Global imization methods for high-dimensional problems. *European Journal of Operational Research*, 119:345–352, 1999.

[93] J. Schroeter, P. Meyer, and S. Parthasarathy. Evaluation of improved articulatory codebooks and codebook acess distance measures. In *ICCASP '92*, pages 393–396, 1992.

[94] J. Schroeter and M. M. Sondhi. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(1):133–150, 1994.

[95] Juergen Schroeter and Mohan Sondhi. *Advances in Speech Signal Processing*, chapter Speech coding based on physiological models of speech production, pages 231–267. Marcel Decker, 1992.

[96] Maria Schuster and et. al. Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition. *International Journal of Pediatric Otorhinolaryngology*, 70(10):1741–1747, 2006.

[97] Alexander Sepulveda, Julian D. Arias, and G. Castellanos-Dominguez. Acoustic-to-articulatory mapping of tongue position for voiced speech signals. In *Advanced Voice Function Assessment International Worshop, AVFA-2009*, 2009.

[98] Alexander Sepúlveda, Germán Castellanos-Dominguez, and Juan Godino-Llorente. Acoustic analysis of the unvoiced stop consonants for detecting hypernasal speech. In *4th International Symposium on Image/Video Communications over fixed and mobile networks, ISIVC-2008*, 2008.

[99] Alexander Sepulveda, Edilson Delgado-trejos, Murillo-Rendón, and Germán Castellanos-Dominguez. Hypernasal speech detection by acoustic analysis of unvoiced plosive consonants. *Revista Tecnológicas*, pages 223–237, Diciembre 2009.

[100] Katsuhiko Shirai and Tetsunori Kobayashi. Estimating articulatory motion from speech wave. *Speech Communication*, 5:159–170, 1986.

[101] Jorge Silva and Shrikanth Narayanan. Discriminative wavelet packet filter bank selection for pattern recognition. *IEEE Transactions on Signal Processing*, 57(5), May 2009.

[102] Thorsten Smit, Friedrich Turckheim, and Robert Mores. Fast and robust formant detection from lp data. *Speech Communication*, 54:893–902, 2012.

[103] V. Sorokin and A. Trushkin. Articulatory-to-acoustic mapping for inverse problem. *Speech Communication*, 19, 1996.

[104] Victor Sorokin, Leonov Alexander, and Alexander Trushkin. Estimation of stability and accuracy of inverse problem solution for the vocal tract. *Speech Communication*, 30:55–74, 2000.

[105] Viktor Sorokin. Determination of vocal tract shape for vowels. *Speech Communication*, 11:71–85, 1992.

[106] Ingmar Steiner, Korin Richmond, Ian Marshall, and Calum Gray. The magnetic resonance imaging subset of the mngu0 articulatory corpus. *J. Acoust. Soc. Am.*, 131(2), 2012.

[107] Kenneth S. Stevens. *Acoustic phonetics*. MIT Press, 2000.

[108] Taiji Suzuki, Masashi Sugiyama, Takafumi Kanamori, and Jun Sese. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, 10(1), 2009.

[109] David Talkin. Speech formant trajectory estimation using dynamic programming with modulated transition costs. *J. Acoustical Society America*, S1:S55–S55, 1987.

[110] Tomoki Toda, Alan Black, and Keiichi Tokuda. Statistical mapping between articulatory movements and acoustic spectrum using gaussian mixture models. *Speech Communication*, 50:215–227, 2008.

[111] Y. Tohkura. A weighted cepstral distance measure for speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(10), October 1987.

[112] Asterios Toutios and Konstantinos Margaritis. Contribution to statistical acoustic-to-EMA mapping. In *16th European Signal Processing Conference (EUSIPCO-2008)*, 2008.

[113] Asterios Toutios, Slim Ouni, and Yves Laprie. Estimating the control parameters of an articulatory model from electromagnetic articulograph data. *J. of the Acoustical*

Society of America, 129(5), 2011.

[114] Robert Tibshirani Trevor Hastie and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2008.

[115] Andrew R. Webb. *Statistical Pattern Recognition*. John Wiley & Sons, 2002.

[116] Q. Xue and et. al. Improvement in tracking of articulatory movements with the x-ray microbeam system. In *Annual International Conference on Engineering in Medicine and Biology Society*.

[117] H. Yang, S. Vuuren, S. Sharma, and H. Hermansky. Relevance of time-frequency features for phonetic and speaker channel classification. *Speech Communication*, 31:35–50, 2000.

[118] Z. Yang, K. Tang, and X. Yao. Large scale evolutionary optimization using cooperative coevolution. *Information Sciences*, 178:2985–2999, 2008.

[119] Stylianou Yannis. *Harmonic plus Noise Models for Speech, Combined with Statistical Models, for Speech and Speaker Modification*. PhD thesis, l'Ecole Nationale Supérieure des Télécommunications, 1996.

[120] Atef Ben Youssef. *Control of talking heads by acoustic-to-articulatory inversion for language learning and rehabilitation*. PhD thesis, l'École Doctorale Electronique, Electrotechnique, Automatique & Traitement du Signal (EEATS), Université de Grenoble, 2011.

[121] Atef Ben Youssef, Thomas Hueber, Pierre Badin, and Gérard Bailly. Toward a multi-speaker visual articulatory feedback system. In *InterSpeech*, 2011.

[122] Le Zhang and Steve Renals. Acoustic-articulatory modeling with the trajectory HMM. *IEEE Signal Processing Letters*, 15:245–248, 2008.

[123] Victor Zue and Ronald Cole. Experiments on spectrogram reading. In *IEEE ICASSP*, pages 116–119, Washington, D.C, 1979.