redin Revista Facultad de Ingeniería, Universidad de Antioquia, No. 77, pp. 162-169, 2015

Subject-independent acoustic-to-articulatory mapping of fricative sounds by using vocal tract length normalization

Mapeo acústico-articulatorio independiente del hablante usando normalización de la longitud del tracto vocal

Alexander Sepúlveda-Sepúlveda[1]*, Germán Castellanos-Domínguez[2], Pedro Gómez-Vilda[3]

[1]Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones, Universidad Industrial de Santander. Carrera 27 Calle 9. A. A. 678. Bucaramanga, Colombia.
[2]Grupo de Procesamiento y Reconocimiento de Señales, Facultad de Ingeniería y Arquitectura, Universidad Nacional de Colombia. Campus la Nubia, km. 7 Vía al Magdalena. A. A. 127. Manizales, Colombia.
[3]Departamento de Arquitectura y Tecnología de Sistemas Informáticos, Facultad de Informática, Universidad Politécnica de Madrid. Campus Montegancedo, C. P. 28660. Madrid, España

ARTICLE INFO
Received May 5, 2015
Accepted October 5, 2015

KEYWORDS
Acoustic-to-articulatory inversion, information measure, speech processing, articulatory phonetics

Inversión acústico-articulatoria, medida de información, procesamiento del habla, fonética articulatoria

ABSTRACT: This paper presents an acoustic-to-articulatory (AtoA) mapping method for tracking the movement of the critical articulators on fricative utterances. The proposed approach applies a vocal tract length normalization process. Subsequently, those acoustic time-frequency features better related to movement of articulators from the statistical perspective are used for AtoA mapping. We test this method on the MOCHA-TIMIT database, which contains signals from an electromagnetic articulograph system. The proposed features were tested on an AtoA mapping system based on Gaussian mixture models, where Pearson correlation coefficient is used to measure the goodness of estimates. Correlation value between the estimates and reference signals shows that subject-independent AtoA mapping with proposed approach yields comparable results to subject-dependent AtoA mapping.

RESUMEN: Este artículo presenta un método de mapeo acústico-a-articulatorio que permite hacer seguimiento del movimiento de articuladores críticos en voces del tipo fricativo. El método propuesto hace uso de aquellas características de tiempo-frecuencia mejor relacionadas desde un punto de vista estadístico con el movimiento de los articuladores, en donde, para el proceso de estimación de características, se aplica normalización por longitud del tracto vocal. El método se prueba sobre la base de datos MOCHA-TIMIT, la cual contiene señales provenientes de un articulógrafo electromagnético. El conjunto de características propuesto se prueba sobre un sistema de mapeo acústico-a-articulatorio basado en modelos de mezclas Gaussianas, en donde la correlación de Pearson se utiliza para medir la bondad de las estimaciones. Se hacen pruebas de una manera independiente del hablante, es decir, los conjuntos de entrenamiento y prueba pertenecen a diferentes hablantes. Como resultado se obtiene que el valor de correlación entre las estimaciones, hechas de manera independiente del hablante, es comparable a la medida de desempeño obtenida desde un punto de vista dependiente del hablante.

1. Introduction

The objective of acoustic-to-articulatory (AtoA) inversion is to obtain articulatory information from the acoustic data contained in the speech signal [1]. It offers new perspectives and interesting applications in the speech processing field. An adequate system for recovering the articulatory configurations might be used in several applications: visual aids in articulatory training tasks for hearing or speech impaired people; computer guided second language learning programs to demonstrate correct and incorrect pronunciation; low-bit rate coding since articulators move relatively slow; and, enhancing representation in speech recognition systems to improve their performance since articulatory parameters represent co-articulatory related phenomenon in a better way.

Although several attempts have been made during more than thirty years, the speech researchers still regard the acoustic-to-articulatory inversion as an open issue [2-4]. Roughly, inversion methods can be divided into two

* Corresponding author: Alexander Sepúlveda Sepúlveda
E-mail: fasepul@uis.edu.co
ISSN 0120-6230
e-ISSN 2422-2844

DOI: 10.17533/udea.redin.n77a19

categories: analysis-by-synthesis approaches and data-driven approaches. Several articulatory inversion methods are based on the analysis-by-synthesis approach, which is a closed-loop optimization procedure that involves the comparison of the spectrum of synthesized speech to the measured speech at consecutive frames, for example [4-8].

On the other hand, nonlinear regression-based approaches require a considerable quantity of parallel acoustic-articulatory data. Fortunately, technologies such as electromagnetic 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 vector quantization with codebooks [9], neural networks [10], Gaussian mixture models [11-13], support vector regression [14, 15] and generalized smoothness criterion [16]. However, such approaches may be unsuccessful if acoustic-articulatory information belonging to the test subject is not included in the training data.

Multi-speaker acoustic-to-articulatory inversion based on hidden Markov models (HMM) have been used in [17]; however it requires a data stream with information about the phonemes present in the speech signal. Additionally, the training of the HMM models in [17] requires several speakers, whereas only one speaker is required in the method proposed in this paper. It is important to note that EMA (ElectroMagnetic Articulograph) data are scarce and expensive. Therefore, using a minimal amount of training data is important. Another subject-independent approach has been proposed in [18, 19], 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.

On the other hand, as discussed in [20], certain articulators (termed critical) play a more significant role during the phone production than others. When one articulator constricts for a phoneme, the others are relatively free to co-articulate (if they do not cause an additional constriction). The work in [13] makes use of the critical articulator concept and relevant time-frequency features. In that work, the proposed representation corresponds to acoustic input features being closely related to the position of critical articulators from the statistical perspective. However, non-critical articulators are free to move. As a result, the statistical association measure may have been affected by the intrinsic movements of these articulators, which are speaker dependent. That is, the influence by the critical articulators on the speech signal is expected to be more consistent between speakers than the influence of non-critical articulators. Therefore, as an alternative to using the whole speech signal, the inversion mapping process by articulatory categories may yield better results.

This paper seeks to show that using VTLN (Vocal Tract Length Normalization) in conjunction with statistically relevant parameters produces effective results for the case of acoustic-to-articulatory inversion in a speaker-independent way. In this paper, the algorithm is tested on fricative phonemes, which show a higher degree of complexity in their acoustic performance over other types of phonemes. The approach requires acoustic-articulatory training data from only one speaker and uses the obtained model to perform articulatory inversion on any arbitrary speaker. The proposed input features are tested in an acoustic-to-articulatory inversion system based on Gaussian mixture models. Obtained results show that the proposed approach achieves good correlation value between estimated and reference articulatory signals.

2. Method

2.1. Database

The present study uses the MOCHA-TIMIT database, which contains a collection of sentences that are designed to provide a set of diverse phonetic utterances. The database includes four data streams recorded concurrently: the acoustic waveform (16 kHz sample rate, with 16 bit precision), laryngograph, electropalatograph, and EMA data. The EMA system is based on the fact that when a spool is 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 [21]. Therefore, when measuring the voltages, the distance can be inferred in relation to a particular point of reference. The location of pellets (spools) is as shown in Figure 1. The two coils at the bridge of the nose and upper incisors, respectively, provide reference points to correct errors produced by head movements.

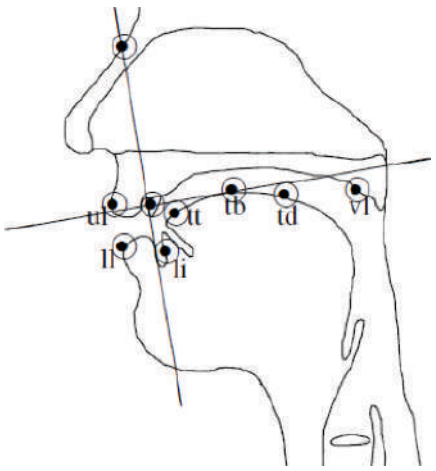


Figure 1 Positions of EMA contacts in the MOCHA-TIMIT database. tt, tongue tip; tb, tongue body; td, tongue dorsum ; li, lower incisors; ll, lower lip; ul, upper lip; vl, velum

Movements of receiver coils attached to the articulators are sampled by the EMA system at 500 Hz. Coils are affixed to the lower incisors (li), upper lip (ul), lower lip(ll), tongue tip (tt), tongue body (tb), tongue dorsum (td), and velum (vl). The MOCHA-TIMIT collection has the acoustic-articulatory data of two speakers: one female (fsew0) and one male (msak0). Although the MOCHA database contains recordings from other speakers, that other data has not been yet adequately corrected and labeled. After filtering with a 8th order Chebyshev Type II low-pass filter of 40 Hz cut-off frequency, the EMA trajectories are resampled from 500 Hz to 100 Hz. Phase response of this filter is approximately linear at the low frequencies of interest; thus, possible distortions caused by filtering process are diminished. Although a sampling frequency of 50 Hz is enough, because muscle contractions typically have bandwidths of up to 15 Hz [9], we selected 100 Hz because it is one of the most common values used in previous works.

Then, a normalization procedure is carried out by using the method explained in [21]. To extract the speech segments corresponding to the fricatives, the labels provided in [22] are used. ll_x , ll_y are critical for phonemes /f, v/; while tt_x and tt_y are critical for the other English fricative phonemes. Thus, two sets of acoustic-articulatory pairs are employed for each speaker. An example of the articulatory data distribution is shown in Figure 2. A detailed explanation about the MOCHA-TIMIT database can be found in [21, 10].

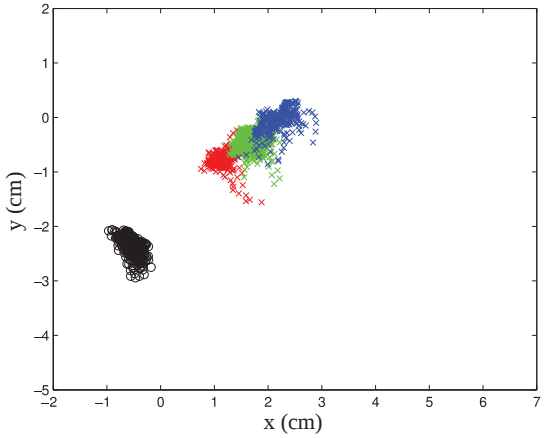


Figure 2 Scatter plot of EMA data corresponding to tongue tip (in colors) and lower lip (in black), which are the critical articulators of fricatives. Colors indicate point of articulation category: dental (red); alveolar (green); and, palatal (blue). The plot is obtained by using the phrase 400th to 460th of msak0 speaker belonging to MOCHA database

2.2. Speech Signal Representation

In order to evaluate the influence of the VTLN process on time-frequency (TF) relevant maps and the contribution of the union of these two techniques in the process of speaker-independent articulatory inversion, two types of representations are generated: the first input set corresponds to those statistically significant energies located in the TF plane. In the second case, we use these same features but previously applying VTLN. The vocal tract length differences between the female (fsew0) and male (msak0) speakers are taken into account during the feature estimation procedure. To diminish their influence, vocal tract length normalization is performed for both speakers by applying the frequency warping functions shown in Figure 3; where, the slopes are obtained from the information reported in [23] (page 42 and tables 4.2-4.3).

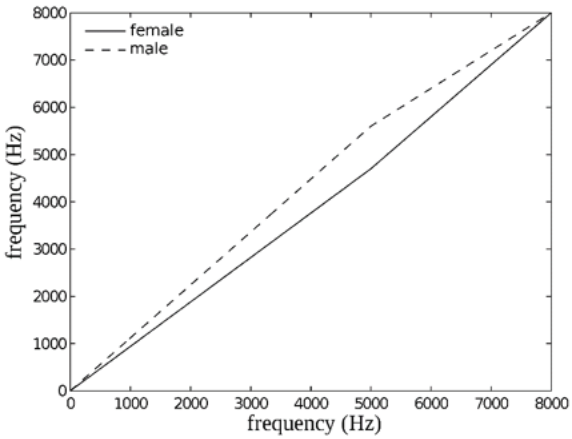


Figure 3 Frequency warping functions used for the vocal tract length normalization process of female and male speakers in MOCHA-TIMIT database. The slopes of the linear functions beginning in the origin are 0.94 and 1.12 for female and male speakers, respectively

In [23], the linear section's length of vocal tract model D is adapted by expanding or compressing it by a factor of K. The adaptation procedure is described by the equation $D = D'(1 - k)$; where, D' stands for reference vocal tract length and D corresponds to vocal tract length of speakers. Note that the value of K reflects vocal tract stretching or compression with respect to the length in Maeda's model [24], thus the value of K for the speaker used to obtain that model is assumed to be 1. It is the reference speaker. The value of K for speakers fsew0 and msak0, in respect to speaker in Maeda's model, found in [23] are: $K = 0.06$ for fsew0 and $K = -0.12$ for msak0. Thus, resulting relative slopes correspond to 0.94 and 1.12 for female and male speakers, respectively. These values are used to obtain the first linear segment of warping functions, which are applied until input frequency reaches 5 KHz, whose value is due to the fact that it is expected that most of the important frequency components lie inside this range.

In this study, frequency splitting is generated with 24 mel filter banks. To carry out the time plane partition, the acoustic speech signal is parameterized using 20 ms width frames and steps of Δt = 10 ms, so a frame rate of 100 Hz is performed. The applied windowing function corresponds

to Hanning window. Acoustic information within a time interval ranging from lower boundary $t\text{-}t_a$ = *t-200 ms* to upper boundary $t+t_b$ = *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 functions shown in Figure 3. Therefore, a total of 50x24=1200 TF atoms representing speech signal behavior at time *t* are obtained. The mel filterbank covers the range of frequencies from 20 Hz to 8000 Hz.

The TF information is represented by the scalar valued logarithmic energy features $x(t + d, f_k) \in R$, where the set $\{f_k: k = 1, \ldots, n_f\}$ appraises the $n_f = 24$ frequency components, being $d \in [t_a, t_b]$ the time-shift variable. A resulting acoustic matrix of log-energy features $X_t \in R^{nt \times nf}$ (with nt *t=* $(t_b - t_a)/10$ *ms*) is attained for each window analysis at the time position *t* of the articulatory configuration $y_t = \{y^m(t) : m = 1, \ldots, n_c\} \in R^{nc \times 1}$, where *m* denotes the m-th channel and $n_c = 4$ is the number of EMA channels employed in this work, (i.e, tt_x, tt_y, ll_x and ll_y).

2.3. Proposed Acoustic Input Features

To estimate the relevant feature set, a statistical measure of association is applied to the TF atoms enclosed within the context window $[t\text{-}t_a, t+t_b]$. Particularly, we use the X^2 information measure $I(x(\cdot), y(\cdot)) \in R$ holding information content with regard to the articulatory trajectory $y^m(t) \in y_t$ of each individual acoustic feature $x(t+d, f_k)$, where $x(t+d, f_k)$ describes the TF-atom at time $t+d$ and frequency f_k in the TF plane X_t.

The X^2 information measure is regarded as the distance between a joint probability distribution, $P_{xy}(\cdot,\cdot)$; and the product of marginal distributions: $P_x(\cdot)$ and $P_y(\cdot)$; respectively [25]. Its estimation is based on the density ratio concept, as shown in Eq. (1). It is explained in detail in [26]:

$$I\big(x(t+d,f_k), y^m(t)\big) = \iint \frac{\left(P_{xy}\big(x(t+d,f_k),\ y^m(t)\big) - P_x\big(x(t+d,f_k)\big)P_y(y^m(t))\right)^2}{P_x(x(t+d,f_k))P_y(y^m(t))}\,dx\,dy \tag{1}$$

The information content of (1) 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 shown in Eq. (2) [25]:

$$I^m_{d,k} = I\big(x(t+d,f_k), y^m(t)\big) = \iint r^m_{d,k}(t)\,dx\,dy \tag{2}$$

where the term $r^m_{d,k} \in R$ in Eq. (2) is determined as shown by Eq. (3):

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

Here, the process generates 1200 statistical association values at each time *t*. A maximum of 2000 pairs $\{x_t, y^m(t)\}$ of EMA-acoustic points are taken for the estimation of relevant TF features. The X^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.

2.4. Acoustic-to-Articulatory Mapping Approach

The current work consists on searching the estimation $\tilde{y}_t$ of the articulatory configuration y_t from the acoustic vector $v_t \in R^{p\times 1}$, comprising *p* selected TF features at the time moment *t; i.e,* $\tilde{y}_t = E\{y_t/v_t\} = \int P(y_t/v_t)\, y_t\, dy_t$. We assume that *y, v* are jointly distributed and that they can be represented in terms of a mixture of Gaussians as shown in (4),

$$P(z_t;\cdot) = \sum_{j=1}^{J} \pi^j\, N\big(z_t; \mu_z^j; \Sigma_z^j\big) \tag{4}$$

where, z_t in Eq. (4) is the joint vector $[v_t^T\ y_t^T]$ and π^j is the weight of the j_{th} mixture component. ^{T} denotes the transpose of the vector. The mean vector μ_z^j and covariance matrix Σ_z^j of the j_{th} mixture component are denoted by Eq. (5) as follows [27],

$$\mu_z^j = \begin{bmatrix}\mu_v^j \\ \mu_y^j\end{bmatrix} \qquad \Sigma_z^j = \begin{bmatrix}\Sigma_{vv}^j & \Sigma_{vy}^j \\ \Sigma_{yv}^j & \Sigma_{yy}^j\end{bmatrix} \tag{5}$$

The conditional probability, see Eq. (6), can also be expressed as a GMM as follows,

$$P\left(y/v\,; \mu_{y/v}^j; \Sigma_{y/v}^j\right) = \sum_{j=1}^{J} \beta_{v_t}^j N\left(y;\ \mu_{y/v}^{j,t}\ ;\ \Sigma_{y/v}^j\right) \tag{6}$$

where the parameter in Eq. (6) corresponding to conditional mean is calculated by Eq. (7),

$$\mu_{y/v}^{j,t} = \mu_y^j + \Sigma_{yv}^j\big(\Sigma_{vv}^j\big)^{-1}\big(v_t - \mu_v^j\big) \tag{7}$$

Similarly, Eq. (8) is the conditional covariance,

$$\Sigma_{y/v}^j = \Sigma_{yy}^j - \Sigma_{yv}^j\big(\Sigma_{vv}^j\big)^{-1}\Sigma_{yv}^j \tag{8}$$

On the other hand, $\beta^j(v_t)$ is computed by using the expression in Eq. (9):

$$\beta^j(v_t) = \frac{\pi^j N(v_t; \mu_v^j; \Sigma_v^j)}{\sum_{i=1}^{J} \pi^i N(v_t; \mu_v^i; \Sigma_v^i)} \tag{9}$$

Lastly, the estimation $\tilde{y}_t$ yields the expression provided by Eq. (10) [11]

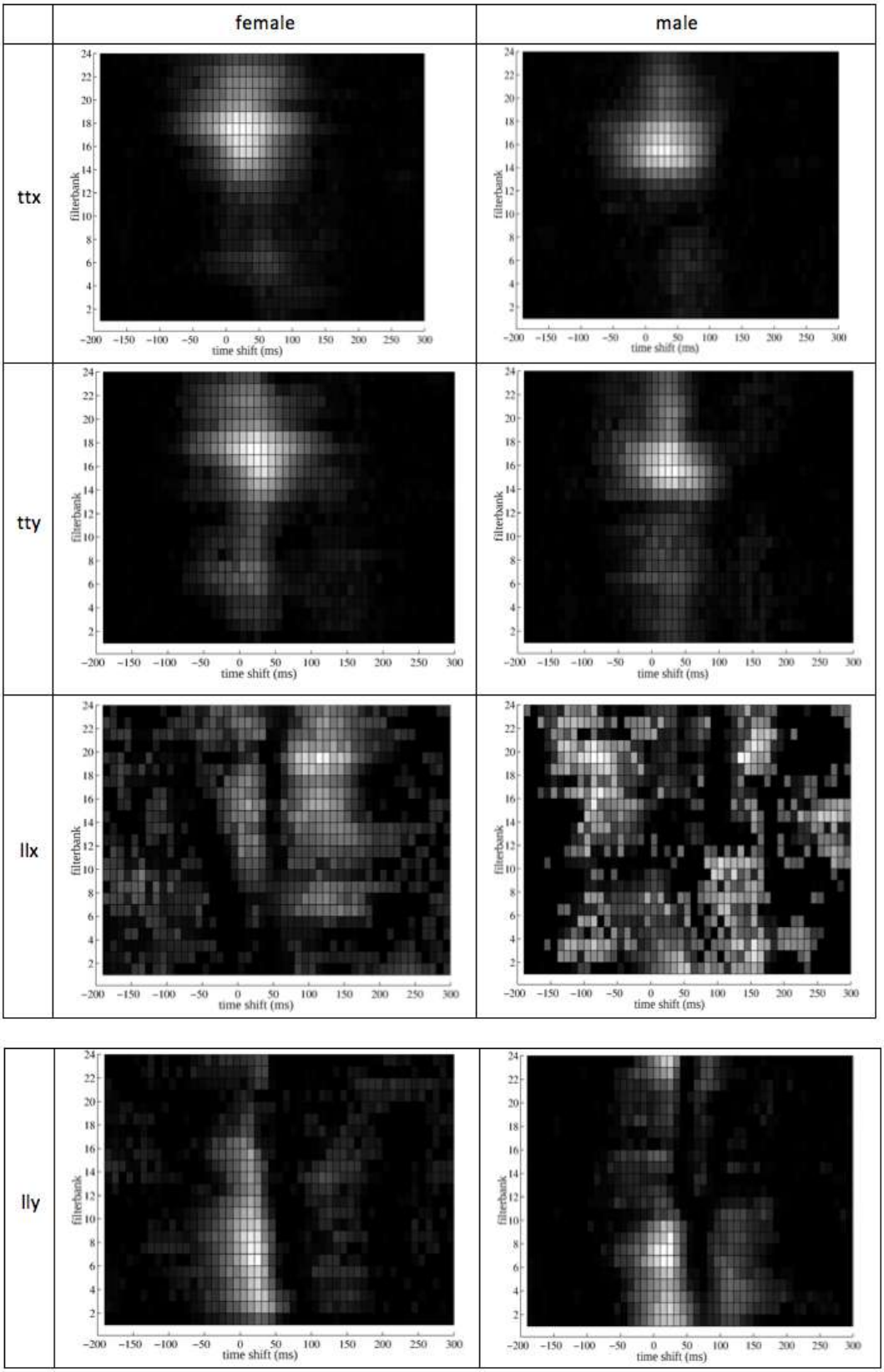


Figure 4 Relevant time-frequency atoms for the critical articulators of the fricative phonemes: tt_x, tt_y, ll_x, ll_y. The maps are obtained after applying VTLN process by using the warping function shown in Figure 3

$$\tilde{y}_t = \sum_{j=1}^{J} \beta^j(v_t)\left(\mu_y^j + \Sigma_{yv}^j\left(\Sigma_{vv}^j\right)^{-1}\left(v_t - \mu_v^j\right)\right) \quad (10)$$

3. Results

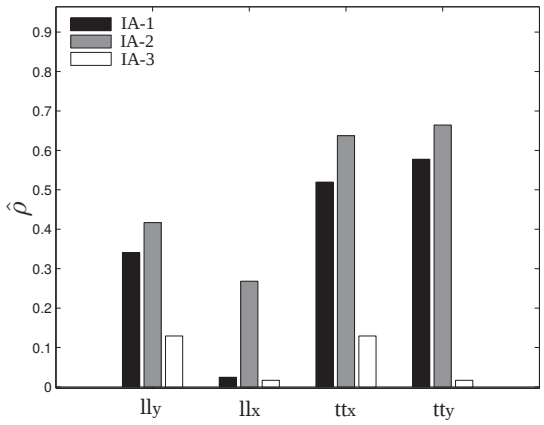
The relevant features are estimated for the female and male speakers of the MOCHA database. As seen in Figure 4 showing the relevant TF atoms corresponding to fricatives, the relevant zones for the channel ll_x are very diffuse. This observation agrees with the distribution of lower lip shown in Figure 2, where it can be seen that the major part of the variance is along the y axis.

In the experiments we consider three approaches to be compared: a) the proposed subject-independent inversion method (noted as IA-1), that makes use of relevant TF features and the VTLN procedure; b) the IA-2 approach corresponding to the conventional subject-dependent method used in previous works [11, 13], 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. Articulatory trajectories are estimated by using Eq. (10).

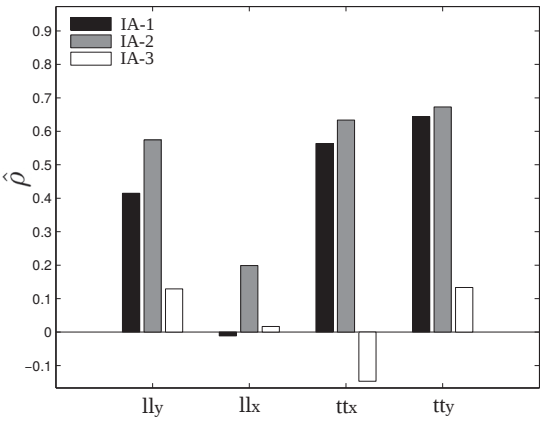
In case of IA-2 and IA-3 approaches, the number of inputs ranges from p = 24 to p = 120 (p = 24; 72, and 120); that is, 1, 3; and 5 frames around current time of analysis were taken into account. The input vector was projected using Principal Component Analysis, where np = 24; 35; 35 components were taken, respectively. When employing the IA-1 approach, the p = 24; 72; and 120 most relevant atoms were used. Then, the np = 24; 35; 35 principal components were extracted to form the input vector for the model given in Section II-D. In all cases, 32 mixtures were used. The model parameters were found by using the expectation maximization algorithm.

Although for the same articulator, the raw EMA data may vary between subjects, the shape of articulatory trajectories is expected to be similar for each one of the phonemes [18]. Therefore, correlation value is used to measure the inversion quality, which is the same evaluation criterion utilized in [19].

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 5 showing the $\hat{\rho}$ values for the msak0 and fsew0 speakers, the proposed method IA-1 offers a better performance with respect to IA-3, and it is comparable to the inversion scheme IA-2. Interestingly enough, in the case of IA-1, testing and training data belong to different speakers. In contrast, IA-2 training and testing data belong to the same speaker. The values obtained in this work using strategy IA-1 are comparable with those obtained in [19] for the same articulators, see Figure 5. In the aforementioned work, average on all EMA channels is 0.53 for female and male speakers.



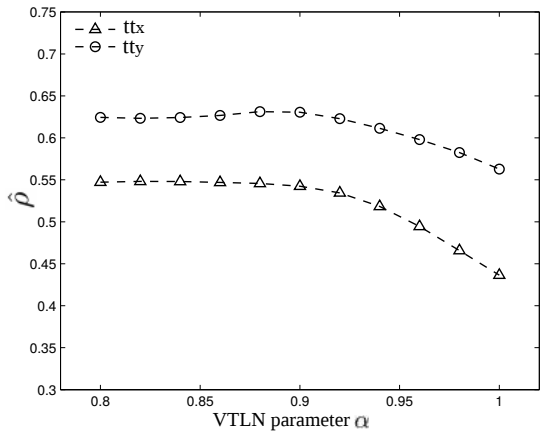
(a)



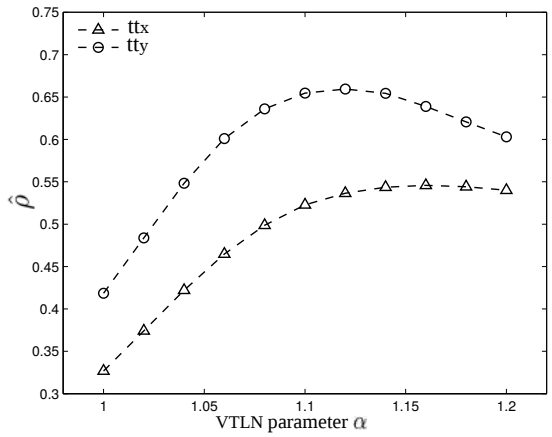
(b)

Figure 5 Correlation average performance $\hat{\rho}$ for speakers fsew0 (a) and msak0 (b) by using the inversion approaches 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

Figure 5 was obtained by using the frequency warping functions of Figure 3 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. This figure was obtained by varying the slope parameter α of the VTLN function and computing performance. In this graph, it can be seen that the performance of the system varies with the value of the slope of the VTLN function. In case of male speaker, it can be observed that the best performance occurs approximately at α = 1.12, whereas the best VTLN parameter is α = 0.9 instead of α = 0.94 for the female speaker. Using α = 0.9 might cause the most relevant features for fsew0 speaker be moved to lower frequency regions.



(a)



(b)

Figure 6 Correlation average performance $\hat{\rho}$ for tongue tip (ttx and tty) of speakers fsew0 and msak0. (a) $\hat{\rho}$ versus VTLN parameter α for fsew0 speaker, (b) $\hat{\rho}$ versus VTLN parameter α for msak0 speaker

4. Conclusion

This work shows that using the adequate input features allows for the inference of critical articulators movement in a subject-independent way. In our case, we utilized time-frequency relevant features after vocal tract normalization. With respect to critical articulators' movement inference, a set of invariant acoustic features for fricatives is obtained. That is, it can be observed that critical articulators tend to consistently influence the acoustics of speech, but not for the case of the non-critical ones. Therefore, better results could be obtained if we focused on inversion of critical articulators. For the sake of having a fully subject-independent inversion strategy, further experiments in a larger database should be carried out; however, the proposed method is promising because the TF features are obtained from one speaker (male/female) and later the method is tested in unseen data coming from a different speaker (female/male).

The existence of inter-articulators correlation phenomenon is well-known. As part of future work, it could be used, with prior estimation of critical articulators trajectories, in order to infer the trajectories of noncritical articulators. In addition, the proposed approach can be utilized for the classification of fricative sounds. Finally, in order to develop a complete inversion system, the application of time-frequency relevant maps should be extended to other phone categories. It should be also used an adequate representation of the vocal-tract. These considerations are part of our future work.

5. References

1. J. Schroeter and M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 133-150, 1994.
2. H. Kjellström and O. Engwall, "Audiovisual-to-articulatory inversion", *Speech Communication*, vol. 51, no. 3, pp. 195-209, 2009.
3. H. Kadri, E. Duflos and P. Preux, "Learning vocal tract variables with multi-task kernels", in *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Prague, Czech Republic, 2011, pp. 2200-2203.
4. S. Panchapagesan and A. 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.*, vol. 129, no. 4, pp. 2144-2162, 2011.
5. K. Shirai and T. Kobayashi, "Estimating articulatory motion from speech wave", *Speech Communication*, vol. 5, no. 2, pp. 159-170, 1986.
6. V. Sorokin, L. Alexander and A. Trushkin, "Estimation of stability and accuracy of inverse problem solution for the vocal tract", *Speech Communication*, vol. 30, pp. 55-74, 2000.
7. S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion", *Journal of Acoustical Society of America*, vol. 118, no. 1, pp. 444-460, 2005.
8. B. Potard, Y. Laprie and S. Ouni, "Incorporation of phonetic constraints in acoustic-to-articulatory inversion", *Journal of Acoustical Society of America*, vol. 123, no. 4, pp. 2310-2323, 2008.
9. J. Hogden *et al.*, "Accurate recovery of articulator positions from acoustics: new conclusions based on human data", *Journal of Acoustical Society of America*, vol. 100, no. 3, pp. 1819-1834, 1996.
10. K. Richmond, S. King, and P. Taylor, "Modelling the uncertainty in recovering articulation from acoustics", *Computer, Speech & Language*, vol. 17, pp. 153-172, 2003.

11. T. Toda, A. Black and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model", *Speech Communication*, vol. 50, no. 3, pp. 215-227, 2008
12. I. Ozbek, M. Hasegawa and M. 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*, vol. 19, no. 5, pp. 1180-1195, 2011.
13. A. Sepulveda, R. Guido and G. Castellanos, "Estimation of relevant time-frequency features using Kendall coefficient for articulator position inference", *Speech Communication*, vol. 55, no. 1, pp. 99-110, 2013.
14. A. Toutios and K. Margaritis, "Contribution to statistical acoustic-to-EMA mapping", in *16th European Signal Processing Conference* (EUSIPCO), Lausanne, Switzerland, 2008, pp. 1-5.
15. A. Sepulveda, J. Arias and G. Castellanos, "Acoustic-to-articulatory mapping of tongue position for voiced speech signals", in *3rd Advanced Voice Function Assessment International Workshop* (AVFA), Madrid, Spain, 2009, pp. 112.
16. P. Ghosh and S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion", *Journal of Acoustical Society of America*, vol. 128, no. 4, pp. 2162-2172, 2010.
17. S. Hiroya and T. Mochida, "Multi-speaker articulatory trajectory formation based on speaker-independent articulatory HMMs", *Speech Communication*, vol. 48, no. 12, pp. 1677-1690, 2006.
18. P. Ghosh and S. Narayanan, "A subject-independent acoustic-to-articulatory inversion", in *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Prague, Czech Republic, 2011, pp. 4624–4627.
19. A. Afshan and P. Ghosh, "Improved subject-independent acoustic-to-articulatory inversion", *Speech Communication*, vol. 66, pp. 1-16, 2015.
20. G. Papcun *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*, vol. 92, no. 2, pp. 688-700, 1992.
21. K. Richmond, "Estimating articulatory parameters from the acoustic speech signal", Ph.D. dissertation, The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK, 2002.
22. P. Jackson and V. Singampalli, "Statistical identification of articulation constraints in the production of speech", *Speech Communication*, vol. 51, no. 8, pp. 695-710, 2009.
23. Z. Al-Bawab, "An analysis-by-synthesis approach to vocal tract modeling for robust speech recognition", Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, USA, 2009.
24. S. Maeda, "A digital simulation method of the vocal-tract system", *Speech Communication*, vol. 1, no. 3-4, pp. 199-229, 1982.
25. T. Suzuki, M. Sugiyama, T. Kanamori and J. Sese, "Mutual information estimation reveals global associations between stimuli and biological processes", *BMC Bioinformatics*, vol. 10, no. 1, 2009.
26. P. Maji, "f-information measures for efficient selection of discriminative genes from microarray data", *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1063-1069, 2009.
27. A. Kain, M. Macon, "Spectral voice conversion for text-to-speech synthesis", in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, USA, 1998, pp. 285-288.