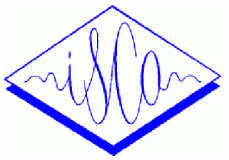

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Inference of Critical Articulator Position for Fricative Consonants

A. Sepulveda[1], *R. Capobianco-Guido*[2], *G. Castellanos-Dominguez*[1]

[1] Machine Learning and Signal Processing Group,
Universidad Nacional de Colombia, Manizales, Colombia
[2]Institute of Physics at São Carlos (IFSC),
University of São Paulo (USP), São Carlos, Brazil
fasepulvedas@unal.edu.co, guido@ieee.org, cgcastellanosd@unal.edu.co

10.21437/Interspeech.2012-222

Abstract

Inversion aims to estimate the articulatory movements which support an acoustic speech signal. Within the acoustic–to–articulatory mapping framework, time frequency atoms had been also employed. The main focus of present work is estimating the relevant acoustic information, in terms of statistical association, for the inference of critical articulators position; in particular, those involved on production of fricatives. The χ^2 information measure is used as the measure statistical dependence. The 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 features are estimated on fricative phones, for which tongue tip and lower lip are 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 shown that relevant features offer potential usefulness on solving the speaker-independent articulatory inversion problem.

Index Terms: Relevant time-frequency features, acoustic-to-articulatory inversion, fricatives, Gaussian mixture models.

1. Introduction

Acoustic-to-articulatory inversion, which is devoted to obtain articulatory information from the acoustic signal [1], offers new perspectives and interesting applications in the speech processing field. To increase the performance of acoustic–to–articulatory mapping systems several aspects are to be considered, among others, the use of an optimal context–window size [3], and to include only those acoustic parameters that are most related to vocal tract mechanism [5]. Since it has been observed that the articulatory information is spread upon several times and frequency ranges of the speech signal, time frequency atoms had been also employed [2, 3]. However the use of compact (non-redundant) and effective time frequency representation poses as an open issue.

To obtain the time frequency atoms more closely related to the articulators movement, measures of statistical dependence might be used. However, this measure is influenced by the critical articulator phenomenon. It is shown in [6] that particular articulators play more significant role to the production of an utterance than others. These articulators are called *critical articulators*. When one articulator constricts for a phoneme, the others are relatively free to coarticulate (assuming that 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. Therefore, if one wishes to establish those most relevant time-frequency features, it is recommended to carry out the analysis by phonetic categories.

This study aims to study the usefulness of time-frequency relevant features in acoustic–to–articulatory inversion tasks. The statistical dependence between the articulatory and the acoustic variables is measured using χ^2 information measure. The benefit of the achieved relevant features is tested in a dependent–speaker acoustic–to–articulatory inversion system based on Gaussian Mixture Models. The potential usefulness of relevant features is tested on an speaker–independent articulatory inversion system.

2. Method

2.1. Database

The present study uses the `MOCHA-TIMIT` database holding a collection of sentences that are designed to provide a set of phonetically diverse utterances. The `MOCHA-TIMIT` database includes four data streams recorded concurrently: the acoustic waveform (16 *kHz* sample rate, with 16 bit precision), laryngograph, electropalatograph, and EMA data. 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 two coils at the bridge of the nose and upper incisors provide

reference points to correct errors produced by head movements. Label files of MOCHA-TIMIT database are used to discard silent segments at the beginning and the end of the utterances [2]. MOCHA-TIMIT database includes the acoustic–articulatory data of two speakers. One is female (fsew0), and the other is male (msak0). 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. Then, the normalization process described in [7] is carried out. For the sake of extracting the speech segments corresponding to fricatives, the labels provided in [8] are used. ll_x and ll_y are critical for phonemes /f, v/; and, tt_x and tt_y are critical for fricative phonemes /θ, ð, s, z, ʃ, ʒ/ [8]. Thus, two sets of acoustic-articulatory pairs are utilized.

2.2. Speech Signal Representation

In this study, frequency splitting is generated with 24 mel filter banks. 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 [2]. Acoustic information within 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.

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 acoustic matrix of 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 $n_c = 4$ is the number of EMA channels employed in present work.

2.3. Estimation of Relevant Features by using 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 because it can be implemented without carrying out an explicit estimation of the joint probability density function [9]. Estimation of the information content by means of the χ^2 measure is written as follows [10]:

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

The expression just described can be estimated based on the density ratio concept. Additional details about the estimation of the density ratio function and the χ^2 information measure can be found in [9, 10].

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 mel filter banks. The process generated 1200 statistical association outcomes for each time t. At maximum of 2000 pairs $\{\boldsymbol{X}_t,\ y^n(t)\}$ of EMA–acoustic points are taken for the estimation of relevant time-frequency features. The χ^2 information measure coefficient between each variable $x(t + d, f_k)$ and articulatory trajectories of the channels corresponding to ll_x, ll_y, tt_x and tt_y is estimated. The resulting points are used to construct the sets of time-frequency relevant features.

2.4. GMM-based Regression

The task at hand consists on searching the estimation $\widetilde{\boldsymbol{y}}_t$ of the articulatory configuration $\boldsymbol{y}_t$ from the acoustic vector $\boldsymbol{v}_t \in \mathbb{R}^{p \times 1}$, comprising p selected TF features at the time moment t, i.e, $\widetilde{\boldsymbol{y}}_t = \boldsymbol{E}\{\boldsymbol{y}|\boldsymbol{v} = \boldsymbol{v}_t\}$. We assume that $\boldsymbol{y}, \boldsymbol{v}$ are jointly distributed. It used the fact that when partitioning the multivariate Gaussian joint density into $P(\boldsymbol{v}, \boldsymbol{y}) = P(\boldsymbol{y}|\boldsymbol{v})P(\boldsymbol{v})$; both $P(\boldsymbol{v}, \boldsymbol{y})$ (conditional probability density function) and $P(\boldsymbol{v})$, they are also multivariate Gaussian. Conditional probability can be represented as $P(\boldsymbol{v}, \boldsymbol{y}) = \sum_{j=1}^{J} \pi^j P(\boldsymbol{y}|\boldsymbol{v}; \cdot)P(\boldsymbol{v}; \cdot)$; where $P(\boldsymbol{y}|\boldsymbol{v}; \cdot)$ can be represented by $P(\boldsymbol{v}; \boldsymbol{\mu}_v^j, \boldsymbol{\Sigma}_v^j) = \sum_{j=1}^{J} \pi^j \mathcal{N}(\boldsymbol{v}; \boldsymbol{\mu}_v^j, \boldsymbol{\Sigma}_v^j)$. The resulting conditional probability can also be expressed as a GMM, as follows: $P(\boldsymbol{y}|\boldsymbol{v}; \boldsymbol{\mu}_{y|v}^j, \boldsymbol{\Sigma}_{y|v}^j) = \sum_{j=1}^{J} \beta^j(\boldsymbol{v}_t)\mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_{y|v}^{j,t}, \boldsymbol{\Sigma}_{y|v}^j)$; where the parameter $\boldsymbol{\mu}_{y|v}^{j,t} = \boldsymbol{\mu}_v^j + \boldsymbol{\Sigma}_{yv}^j(\boldsymbol{\Sigma}_v^j)^{-1}(\boldsymbol{v}_t - \boldsymbol{\mu}_v^j)$ is the conditional mean whereas $\boldsymbol{\Sigma}_{y|v}^j = \boldsymbol{\Sigma}_v^j - \boldsymbol{\Sigma}_{yv}^j(\boldsymbol{\Sigma}_v^j)^{-1}\boldsymbol{\Sigma}_{yv}^j$ is the conditional covariance. $\beta^j(\boldsymbol{v}_t)$ is computed by using the following expression:

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

Lastly, estimation $\widetilde{\boldsymbol{y}}_t$, expressed by expectation over $P(\boldsymbol{y}|\boldsymbol{v})$, yields:

$$\widetilde{\boldsymbol{y}}_t = \boldsymbol{E}\{P(\boldsymbol{y}|\boldsymbol{v})\} = \sum_{j=1}^{J} \beta^j(\boldsymbol{v}_t)(\boldsymbol{\mu}_v^j + \boldsymbol{\Sigma}_{yv}^j(\boldsymbol{\Sigma}_{vv}^j)^{-1}(\boldsymbol{v}_t - \boldsymbol{\mu}_v^j)) \tag{2}$$

3. Results

3.1. Acoustic-to-articulatory mapping using time-frequency relevant features

The TF planes of relevant features are constructed using 10 ms shift rate, the same used in [2, 3, 11]. The relevant time–frequency atoms in case of fricatives /θ, ð, s, z, ʃ, ʒ/ for male a female speaker are shown in figure 1. Even though a frame step of 18 ms is utilized in [4], 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 (2). 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.

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 (2). It can be observed that the performance of acoustic–to–articulatory mapping system increases for the articulators involved in the production of fricatives.

3.2. Time-frequency relevant features for subject independent inversion

Acoustic–to–articulatory mapping is usually done in a speaker-dependent way; therefore, the inversion procedure may not work well when testing the inversion procedure with a speaker whose data are not included in the training set. Present section tests the usefulness of relevant time–frequency features on a speaker–independent framework. The proposed scheme requires acoustic-articulatory training data from only one speaker and uses

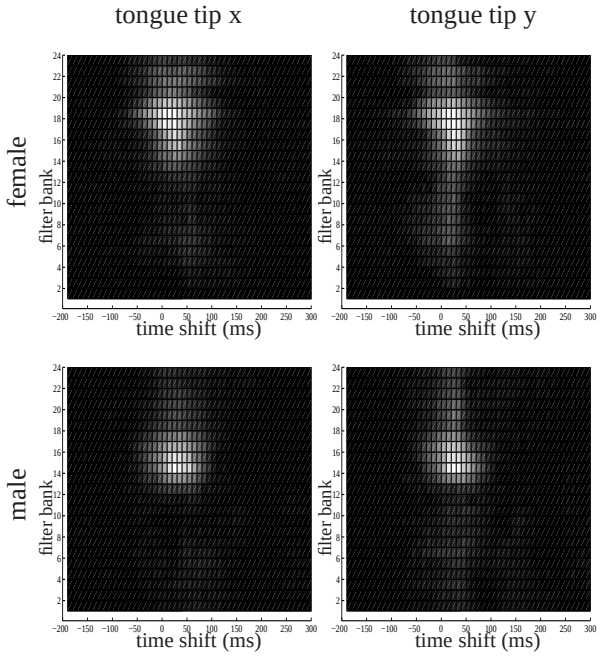

Figure 1: Relevant time–frequency atoms for the critical articulators of the fricative phonemes /θ, ð, s, z, ʃ, ʒ/. (tt_x and tt_y).

the obtained model to perform articulatory inversion on any arbitrary speaker.

Note that there are only two subjects in the MOCHA corpus. The relevant features are estimated for the female speaker of the MOCHA database. The relevant atoms corresponding to fricatives /θ, ð, s, z, ʃ, ʒ/ are shown in figure 1. It is required to take into account the difference in length of the vocal tract for the female speaker (fsew0) and male speaker (msak0). In order to diminish its influence, linear vocal tract normalization is performed for the speaker whose data are not part of the training set. The normalization parameter is calculated by using the vocal tract lengths of speakers fsew0 and msak0 provided in [12].

For the sake of comparison, we have considered the subject independent inversion using GMM. It is equivalent to the conventional method refered in present work, and commonly used recently, except that the training data are obtained from one subject while the other subject's data are used for testing (Inversion scheme 1, IS1). The speaker–independent inversion scheme using relevant features and vocal tract normalization, the scheme proposed in present section, is referred as IS2 (inversion scheme 2). Ranges of the raw EMA data for the same articulator may be different between subjects, but the shape of articulatory trajectories are expected to be similar when two subjects utter the same phoneme [13]. Therefore, correlation value is used to measure the inversion quality. The correlation results for the msak0

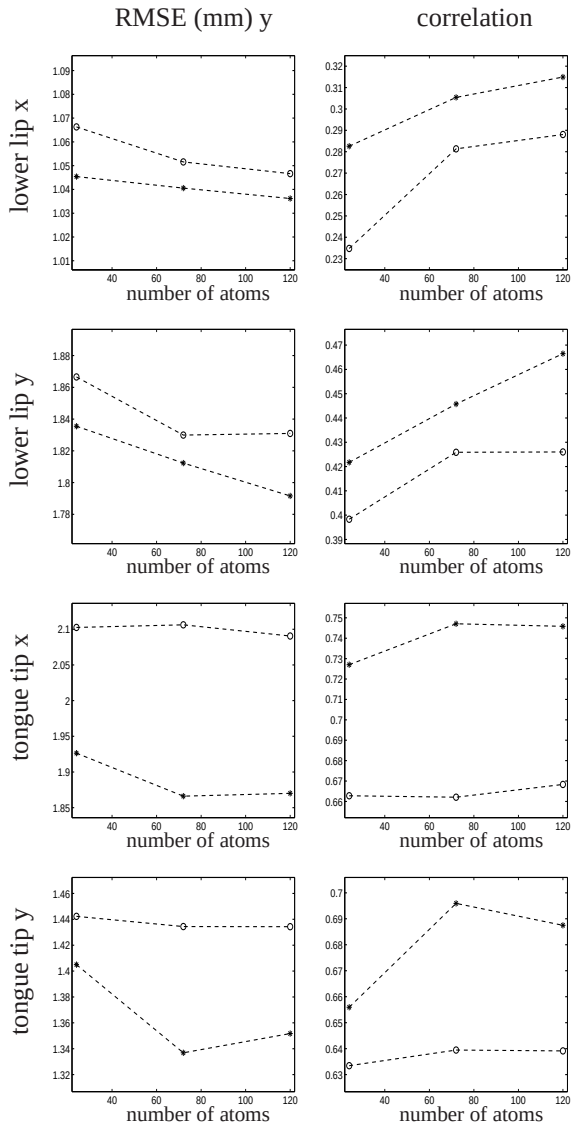

Figure 2: Performance in terms of RMSE and correlation using conventional method (noted with ○) and using relevant time–frequency atoms (noted with ∗) for the critical articulators of fricative consonants.

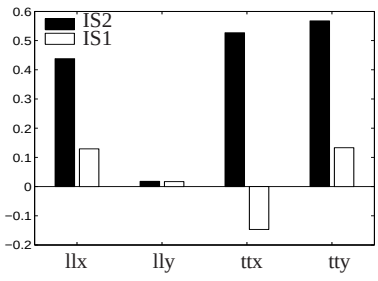

Figure 3: Bar diagram of the correlation value obtained using IS1 and IS2 methods for the channels ll_x, ll_y, tt_x and tt_x

speaker are shown in figure 3. It can be observed that IS2 method offers a noticeable better performance, which is comparable to the one found in [13].

4. Conclusions

The proposed method, which obtains a set of relevant time–frequency components closely related to the articulatory positions on fricatives, is shown to be suitable for improving the performance of acoustic–to–articulatory inversion systems; particularly those based on Gaussian mixture models. Furthermore, the same features show potential usefulness in solving the speaker-independent articulatory inversion problem.

5. References

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

[2] Richmond, Korin, and King, Simon, and Taylor, Paul, "Modelling the Uncertainty in Recovering Articulation from Acoustics", Computer, Speech & Language, 17:153–172, 2003.

[3] Toda, Tomoki, and Black, Alan, and Tokuda,, Keiichi, "Statistical Mapping between Articulatory Movements and Acoustic Spectrum using Gaussian Mixture Models", Speech Communication, 50:215–227, 2008.

[4] Ozbek, Yucel, and Hasegawa–Johnson, Mark, and Demirekler, Mubeccel, "Estimation of Articulatory Trajectories Based on Gaussian Mixture Model (GMM) with Audio-Visual Information Fusion and Dynamic Kalman Smoothing", IEEE Trans. on Audio, Speech, and Language Proc., 19(11):July, 2008.

[5] Qin, Chao, and Carreira-Perpiñan Miguel A., "A Comparison of Acoustic Features for Articulatory Inversion", InterSpeech 2007.

[6] Papcun, George, 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, 92(2):688-700, 1992.

[7] Richmond, Korin, "Articulatory feature recognition from the acoustic speech signal", PhD. Thesis, University of Edinburgh, Edinburgh, 2001.

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

[9] Suzuki, Taiji, and et. al., "Mutual Information Estimation Reveals Global Associations between Stimuli and Biological Processes", BMC Bioinformatics, 10(1), 2009.

[10] Maji, Pradipta, "f-Information Measures for Efficient Selection of Discriminative Genes From Microarray Data", IEEE Trans. on Biomedical Engineering, 56(4):1063-1069, 2009.

[11] Al–Moubayed, Samer, and Ananthakrishnan G., "Acoustic–to–Articulatory Inversion based on Local Regression", InterSpeech–2010, 2010.

[12] Al Bawab, Ziad, "An analysis-by-synthesis approach to vocal tract modelling for robust speech recognition", PhD. Thesis, Carnegie Mellon University, Pittsburgh, 2009.

[13] Prasanta Kumar Ghosh and Shrikanth Narayanan, "A subject-independent acoustic-to-articulatory inversion", In Proceedings of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP), 2011.