TIME-FREQUENCY RELEVANT FEATURES FOR CRITICAL ARTICULATORS MOVEMENT INFERENCE

Alexander Sepulveda
Escuela Colombiana de Carreras Industriales
Facultad de Ingeniería
Kra. 19 No. 49-20, Bogotá, Colombia

Germán Castellanos-Domínguez
Universidad Nacional de Colombia, Sede Manizales
Signal Processing and Recognition Group
Km. 9 vía Al Aeropuerto, Manizales, Colombia

Rodrigo Capobianco Guido
University of São Paulo (USP)
Institute of Physics at São Carlos (IFSC)
São Carlos, SP, Brazil

ABSTRACT

This paper presents a method to study the distribution of the articulatory information on time–frequency representation calculated from the acoustic 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 critical articulator positions. The rank correlation Kendall coefficient is used as the relevance 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 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.

Index Terms— Acoustic–to–articulatory inversion, relevant time–frequency features, gaussian mixture models, wavelet packet transform, articulatory phonetics.

1. INTRODUCTION

Speech gestures are planned movements in a coordinated sequence, which are controlled by intrinsic and extrinsic muscles, and whose actions are relatively slow and overlapping. This circumstances causes the human speech articulators (jaw, tongue, lips, etc.) to have limited freedom of movement and to be interrelated and ruled by inertia. As a consequence, in the 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 neighbouring segments [1]. That is, the information about a phoneme is not localized just on a single phoneme's region, but is spread over a substantial segment of the speech signal. Recent experiments support this affirmation, specially in [2, 3], the use of the mutual information applied to estimation of the distribution of the phonetic information in frequency as well as in time is discussed. On the other hand, the distribution of the articulatory information on the acoustic speech signal is also important; however, its estimation remains unresolved issue. 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, whose main goal is to infer the articulators position based on the information immersed in the acoustic speech signal [4, 5].

It is shown in [6] that certain articulators play more significant role to the production of a given phone than others. These articulators are called *critical articulators*. 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.

This study aims to estimate the influence zones of the critical articulators movement of speakers over time and frequency domains of speech signals. For this purpose, statistical dependence between the articulatory and the acoustic variables is measured by using the Kendall τ coefficient, which is a measure implemented by robust and simple algorithms. As a result, the maps of most relevant zones in time and in frequency for critical articulators movement estimation of the MOCHA–TIMIT speakers database are achieved. The benefit

This work was supported mainly by Administrative Department of Science, Technology and Innovation of Colombia (COLCIENCIAS).

of the achieved relevant zone maps is tested in an acoustic–to–articulatory regression system based on GMMs. It must be noted that the acoustic speech signal is represented using the wavelet packet transform (WPT) that allows a flexible choice of time–frequency bands and can be efficiently implemented, as shown in [7, 8].

2. METHOD

2.1. Speech signal representation

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 [8]. In addition, they offer an alternative for detecting sudden bursts in slowly varying signals [9], which is a phenomenon observed in stop consonants. Through this study, the acoustic speech signal is parameterized using the wavelet packet transform [10], whereas the articulatory information is represented by trajectories extracted from an EMA system that measures the movement of specific points of speech articulators, where each channel becomes a single articulatory dimension.

In this study, frequency splitting of the time–frequency (TF) plane is generated with the Wavelet–Packet Transform (WPT) having frequency bands spacing similar to the Mel scale, as proposed in [11]. WPT of the speech frame at time $t+d$, is computed by using Daubechies compactly supported wavelets with six vanishing moments. The energy of each frequency band that results from the sum of square values of the coefficients contained in the WPT–related nodes is calculated; then, logarithmic operation is performed over attained set of energy values. 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 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 stands for the number of EMA channels.

2.2. Kendall τ 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 [12]:

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

Both terms of $\tau \in [-1, 1]$, in (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 $a_{ij} \in [-1, 1]$ is defined in equation (2) as:

$$a_{ij} = \operatorname{sgn}(x_i(t+d, f_k) - y_i^m(t))\,(x_j(t+d, f_k) - y_j^n(t)) \quad (2)$$

where $\operatorname{sgn}(\cdot)$ stands for the signum function. Then, the value of $\tau_{d,k}^m = \boldsymbol{E}\{a_{ij}\}$ 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:

$$\tau_{d,k}^m = \sum_{1 \le i < j \le N} \sum \frac{a_{ij}}{\binom{N}{2}} \quad (3)$$

2.3. GMM 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, $\widetilde{\boldsymbol{y}}_t = \boldsymbol{E}\{\boldsymbol{y}|\boldsymbol{v} = \boldsymbol{v}_t\} = \int P(y_t|v = v_t) y_t dy_t$. We assume that $\boldsymbol{y}, \boldsymbol{v}$ are jointly distributed and that can be represented in terms of a mixture of gaussians by $P(\boldsymbol{z_t}; \cdot) = \sum_{j=1}^{J} \pi^j \mathcal{N}(\boldsymbol{z_t}; \boldsymbol{\mu}_z^j, \boldsymbol{\Sigma}_z^j)$; where, $\boldsymbol{z_t}$ is the joint vector $\boldsymbol{z_t} = [\boldsymbol{v}_t^\intercal, \boldsymbol{y}_t^\intercal]$ and π^j is the weight of the j_{th} mixture component. $^\intercal$ denotes the transpose of the vector. The mean vector $\boldsymbol{\mu}_z^j$ and covariance matrix $\boldsymbol{\Sigma}_z^j$ of the j_{ht} mixture component are written as [13],

$$\boldsymbol{\mu}_z^j = \begin{bmatrix} \boldsymbol{\mu}_v^j \\ \boldsymbol{\mu}_y^j \end{bmatrix} \qquad \boldsymbol{\Sigma}_z^j = \begin{bmatrix} \boldsymbol{\Sigma}_{vv}^j & \boldsymbol{\Sigma}_{vy}^j \\ \boldsymbol{\Sigma}_{yv}^j & \boldsymbol{\Sigma}_{yy}^j \end{bmatrix} \quad (4)$$

The conditional probability can also be expressed as a GMM, as follows:

$$P(\boldsymbol{y}|\boldsymbol{v}; \boldsymbol{\mu}_{y|v}^j, \boldsymbol{\Sigma}_{y|v}^j) = \sum_{j=1}^{J} \beta^j(\boldsymbol{v}_t) \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_{y|v}^{j,t}, \boldsymbol{\Sigma}_{y|v}^j) \quad (5)$$

where the parameter $\boldsymbol{\mu}_{y|v}^{j,t} = \boldsymbol{\mu}_y^j + \boldsymbol{\Sigma}_{yv}^j (\boldsymbol{\Sigma}_v^j)^{-1} (\boldsymbol{v}_t - \boldsymbol{\mu}_v^j)$ is the conditional mean whereas $\boldsymbol{\Sigma}_{y|v}^j = \boldsymbol{\Sigma}_{yy}^j - \boldsymbol{\Sigma}_{yv}^j (\boldsymbol{\Sigma}_{vv}^j)^{-1} \boldsymbol{\Sigma}_{yv}^j$ is the conditional covariance. $\beta^j(\boldsymbol{v}_t)$ is computed by using the following expression:

$$\beta^j(\boldsymbol{v}_t) = \frac{\pi^j \mathcal{N}(\boldsymbol{v}_t; \boldsymbol{\mu}_v^j, \boldsymbol{\Sigma}_v^j)}{\sum_{i=1}^{J} \pi^i \mathcal{N}(\boldsymbol{v}_t; \boldsymbol{\mu}_v^i, \boldsymbol{\Sigma}_v^i)} \quad (6)$$

Lastly, the estimation $\widetilde{\boldsymbol{y}}_t$ yields $\widetilde{\boldsymbol{y}}_t = \sum_{j=1}^{J} \beta^j(\boldsymbol{v}_t)(\boldsymbol{\mu}_v^j + \boldsymbol{\Sigma}_{yv}^j(\boldsymbol{\Sigma}_{vv}^j)^{-1}(\boldsymbol{v}_t - \boldsymbol{\mu}_v^j))$.

3. RESULTS

3.1. Dataset

The present study uses the MOCHA-TIMIT database holding a collection of sentences that are designed to provide a set of phonetically diverse utterances [14]. The MOCHA-TIMIT database includes the acoustic waveform (16 kHz sample rate) as well as 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 (v). 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 [15]. 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 low–pass filtering process. Then, the normalization process described in [16] is carried out.

The phones for which a given articulator is critical are segmented by using the corresponding MOCHA database labels, which were corrected in [1]. In order to establish correspondence between articulators and phonemes for which the given articulator is critical, IPA descriptors are utilized. They are described as follows: ul_y : /p, b, m/; ll_x : /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/; and, v_x: /m, n, ŋ/.

3.2. Relevant Time–Frequency Maps

For the sake of constructing the maps of relevant features, the Kendall τ coefficient between each variable $x(t+d, f_k)$ and articulatory trajectories of ul_y, ll_x, tt_x, tt_y, td_y and v_x is obtained. At maximum of 5000 pairs $\{\boldsymbol{X}_t,\ y^n(t)\}$ of EMA–acoustic points for male a female speakers are taken. The statistical measure of association, Kendall τ coefficient, 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 2.1. The process generated 1200 statistical association outcomes for each time t. The maps are constructed using 10 ms shift rate, the same used in [15, 13, 17, 18]. Relevant maps are shown in Figures 1 and 2 for speakers fsew0 and msak0, respectively.

Some similarities in shape can be observed between the maps of female and male speaker, in particular for the case of upper lip y, tongue tip x, tongue tip y, tongue dorsum y and velum x; however, the similarities are conditioned by frequency ranges. By contrast, clear differences can be appreciated in case of lower lip x. We offer no explanations to this fact.

3.3. GMM acoustic–to–articulatory mapping

In order to show the usefulness of relevant maps, GMM acoustic–to–articulatory mapping is performed using conventional method, like in [13], as well as using relevant features privided by relevant maps. In case of conventional method, 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. 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 for the model in (**??**). In all cases 32 mixtures are used. The model parameters are found by using the expectation maximization (EM) algorithm [19].

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 [13]. 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 (3.2). 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 performance is measured by using the root mean square error and the Pearson's correlation coefficient. 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 (ul_y, ll_x, tt_x, tt_y, td_y and v_x,). The results are shown in Figure (3). It can be observed that the performance of acoustic–to–articulatory mapping system increases for the articulators just mentioned, except for tt_y.

4. DISCUSSION AND CONCLUSIONS

The proposed method, which obtains a set of relevant time–frequency components closely related to the articulatory positions, is shown to be suitable for improving the performance of acoustic–to–articulatory mapping systems, particularly those based on Gaussian mixture models, as oberved in Figure 3. Moreover, the relevant maps provide a more deeper understanding into the relationship between the articulatory and acoustical phenomena.

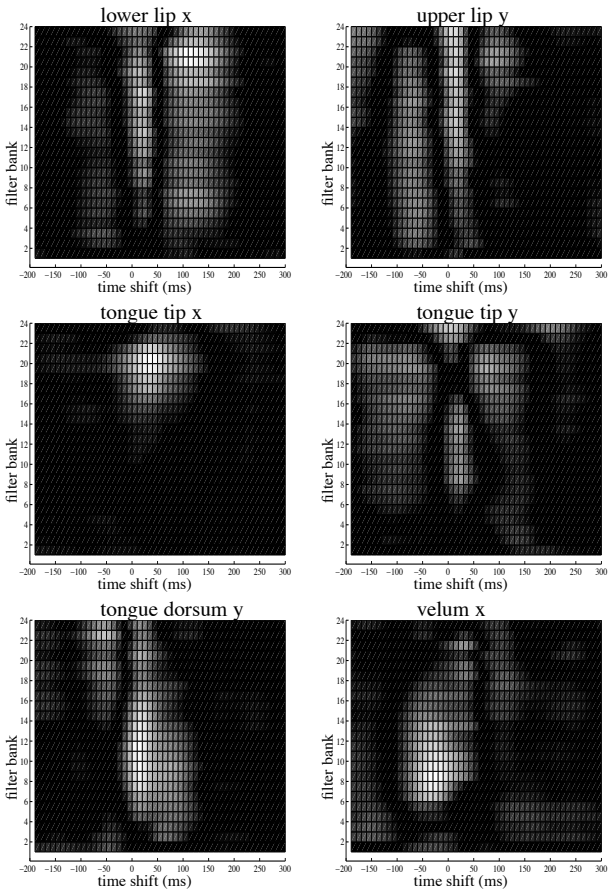

Fig. 1. Relevant time–frequency atoms for female speaker

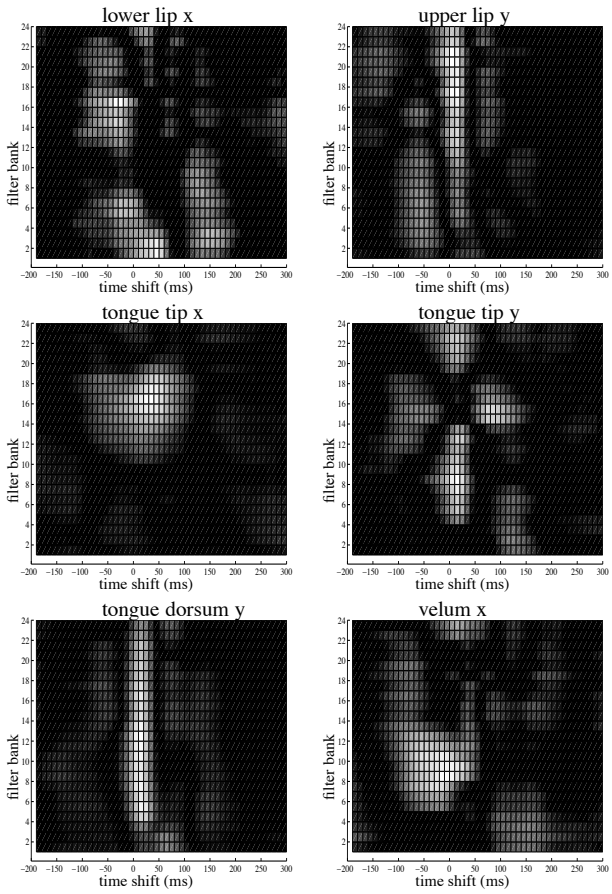

Fig. 2. Relevant time–frequency atoms for male speaker

From the estimated relevance maps, see Figures (1, 2), 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. Yet, channel corresponding to velum x is the exception. 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 to female speaker, but preserving the similarities in shape; particularly, when modelling tongue tip x and tongue tip y, see Figures (1, 2). Observing same figures, additional similarities can be appreciated between the relevant maps of the female speaker and the male speaker; though, not for all articulators.

The authors suggest applying this proposed method 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. In addition, further tests should be performed in order to adapt present method to acoustic–to–articulatory mapping systems to later on compare it with other state–of–the—art methods.

5. REFERENCES

[1] Philip Jackson and Veena Singampalli, "Statistical identifcation of articulation constraints in the production of speech," *Speech Communication*, vol. 51, no. 8, 2009.

[2] H. Yang and et. al., "Relevance of time-frequency features for phonetic and speaker channel classification," *Speech Communication*, vol. 31, pp. 35–50, 2000.

[3] Mark Hasegawa-Johnson, "Time–frequency distribution of partial phonetic information measured using mutual information," in *InterSpeech*, 2000, pp. 133–136.

[4] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 133–150, 1994.

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

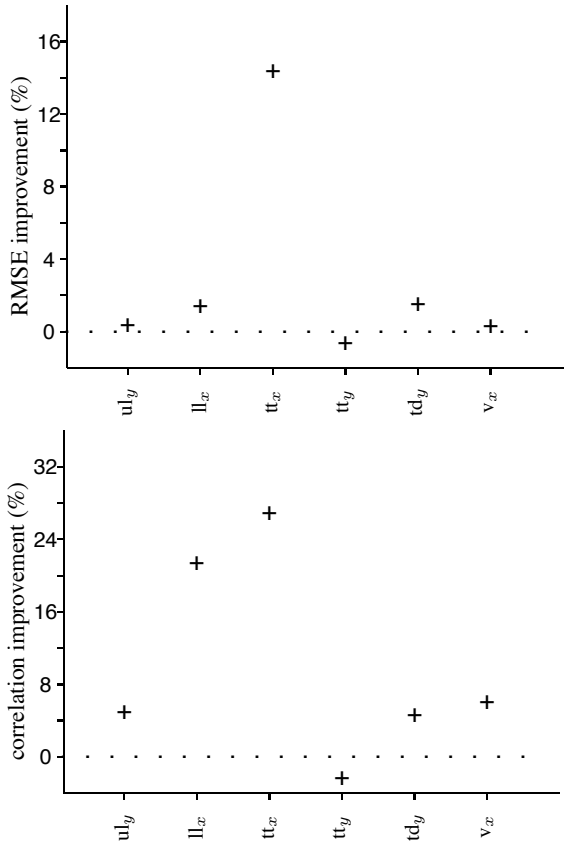

Fig. 3. Percentage improvement of RMSE and correlation using relevant time–frequency atoms based method in respect to using the conventional method for the articulators ul_y, ll_x, tt_x, tt_y, td_y and v_x.

[6] 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*, vol. 2, 1992.

[7] Ghinwa Choueiter and James Glass, "An implementation of rational wavelets and filter design for phonetic classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 3, March 2007.

[8] Jorge Silva and Shrikanth Narayanan, "Discriminative wavelet packet filter bank selection for pattern recognition," *IEEE Transactions on Signal Processing*, vol. 57, May 2009.

[9] Paul S. Addison, *The Illustrated Wavelet Transform*, Institute of Physics Publishing, 2002.

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

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

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

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

[14] Alan Wrench, "The MOCHA-TIMIT articulatory database," Tech. Rep., Queen Margaret University College, www.cstr.ed.ac.uk/research/projects/artic/mocha.html, 1999.

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

[16] Korin Richmond, *Articulatory feature recognition from the acoustic speech signal*, Ph.D. thesis, University of Edinburgh, Edinburgh, 2001.

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

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

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