Formant-controlled HMM-based Speech Synthesis

Ming Lei[1], *Junichi Yamagishi*[2], *Korin Richmond*[2], *Zhen-Hua Ling*[1], *Simon King*[2], *Li-Rong Dai*[1]

[1]iFLYTEK Speech Lab, University of Science and Technology of China, Hefei, China
[2]CSTR, University of Edinburgh, United Kingdom

leiming@mail.ustc.edu.cn

Abstract

This paper proposes a novel framework that enables us to manipulate and control formants in HMM-based speech synthesis. In this framework, the dependency between formants and spectral features is modelled by piecewise linear transforms; formant parameters are effectively mapped by these to the means of Gaussian distributions over the spectral synthesis parameters. The spectral envelope features generated under the influence of formants in this way may then be passed to high-quality vocoders to generate the speech waveform. This provides two major advantages over conventional frameworks. First, we can achieve spectral modification by changing formants only in those parts where we want control, whereas the user must specify all formants manually in conventional formant synthesisers (e.g. Klatt). Second, this can produce high-quality speech. Our results show the proposed method can control vowels in the synthesized speech by manipulating $F1$ and $F2$ without any degradation in synthesis quality.

Index Terms: speech synthesis, hidden Markov model, formant, controllability

1. Introduction

In recent years, hidden Markov model (HMM) based speech synthesis has become a mainstream method, offering high flexibility and naturalness [1]. This method consists of two stages: in the training stage, models are trained using acoustic features, such as spectrum, F0 and duration; in the synthesis stage, acoustic features are predicted by the maximum likelihood parameter generation algorithm, and then sent to a vocoder to construct the waveform. In the statistical parametric speech synthesis method, we can easily change speech characteristics using adaptation and interpolation of model parameters. As this is an automatic, data-driven approach, it is also scalable to very large amounts of data.

However, current systems do not enable structured prior knowledge of speech production or perception to be incorporated in a straightforward way. There are two main reasons for this. First, the acoustic features that are currently used are limited to those required to drive vocoders, such as spectral features, whereas we may wish to introduce prior knowledge in a more structured way, for example in terms of information about the position and dynamics of speech articulators or formants. Second, current approaches largely do not explicitly model the relationships between different levels of acoustic representation. To introduce and utilise such structured prior knowledge in speech synthesis, we need to statistically model not only how speech *sounds* but *how it is produced*.

With this in mind, we have previously developed a two-layer time-series statistical model and have applied it to the joint modelling of spectral features and articulatory features, including tongue movements captured using electromagnetic articulography (EMA) [2]. Although the structural dependency between these sets of features is still approximated by a piecewise linear regression (using a similar technique to speaker adaptation), this model provides significant benefits to statistical speech synthesis: the synthetic speech generated from such a statistical model can be controlled via articulation. More specifically, the generation of acoustic features is not only decided by the acoustic models corresponding to the contextual information, but is also influenced by the concurrent articulatory features. This provides the possibility to control the generation of acoustic features by manipulating those articulatory features. Note that the baseline articulatory parameters are automatically generated in the usual way by optimising a likelihood function. Thereafter, we achieve spectral modification by changing only those parts where we want control. This is an important advantage and should not be confused with conventional articulatory synthesisers, in which the user must specify all articulatory parameters manually.

Training the system in [2] requires EMA data to be recorded in parallel with acoustic data. Since recording EMA data is time-consuming and requires special expertise, it would be more convenient if we could control characteristics of synthesised speech based on phonetic or speech production knowledge without such specialist data acquisition. Therefore, we focus on formants in this paper. Formants too are a meaningful representation with which to characterise speech, especially with respect to speech perception, and we can drawn upon significant existing knowledge about them. Moreover, we can easily calculate formants from the acoustic speech signal alone. Therefore, in this paper we develop a formant-controllable HMM-based speech synthesis system by simply substituting the articulatory layer of [2] with a formant layer, and then investigate how well we can manipulate formants to modify vowels. We hypothesize that all the advantages in controllability that conventional formant synthesisers have (e.g. the Klatt model [3]) can be achieved, while at the same time still producing high-quality speech, since the speech waveform can be generated not from a formant synthesiser but from a high-quality vocoder. This synthesiser could prove very useful in other fields, such as speech perception and phonetics research, where the Klatt model is currently the standard tool.

2. Method

In standard HMM-based speech synthesis, mel-cepstra or line spectral pairs (LSPs) are adopted as spectral features. As resonances of the human vocal tract, formants can characterise speech in a meaningful way. For example, $F1$ (Fj denotes jth formant central frequency) is related to vowel openness,

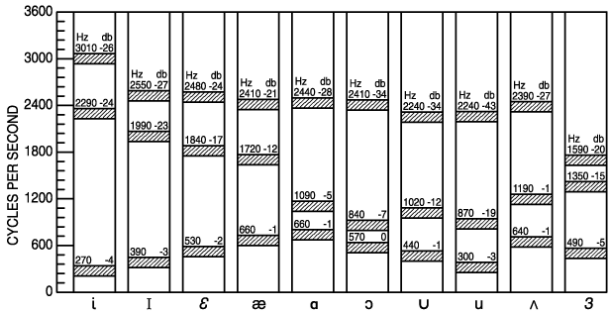

Figure 1: Mean formant frequencies and relative amplitudes for 33 male speakers, for English vowels in a /h-d/ context. This figure was taken from [4].

and $F2$ corresponds to frontness, which are crucial characteristics of vowels. Fig. 1, taken from [4], shows how mean formant frequencies vary for different vowels. Some related work on using formants in speech synthesis has previously been described [5, 6]. The advantages of the method we propose here are: 1) formants affect only statistical models trained on the spectral envelop features (such as LSPs); and 2) speech is hence generated not from formant synthesisers but from high-quality vocoders (such as the STRAIGHT vocoder [7]) using the spectral envelop features generated from the statistical models which are affected by formants.

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be formant and spectral observations respectively, with static and dynamic components, i.e. $\boldsymbol{X} = \boldsymbol{W_X} \boldsymbol{X}_s$ and $\boldsymbol{Y} = \boldsymbol{W_Y} \boldsymbol{Y}_s$, where $\boldsymbol{W_X}$ and $\boldsymbol{W_Y}$ are given in [8]. We define the length of the observations as N, and the parameters of our statistical model as λ. The model used is the same as [2] and is shown in Fig. 2. The likelihood function of their joint distribution is given by

$$\begin{aligned} P(\boldsymbol{X}, \boldsymbol{Y}|\lambda) =& \Sigma_{\forall \boldsymbol{q}} P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{q}|\lambda) \\ =& \Sigma_{\forall \boldsymbol{q}} \pi_{q_0} \prod_{t=1}^{N} a_{q_{t-1} q_t} b_{q_t}(\boldsymbol{x}_t, \boldsymbol{y}_t) \end{aligned} \quad (1)$$

where

$$b_j(\boldsymbol{x}_j, \boldsymbol{y}_j) = b_j(\boldsymbol{y}_j|\boldsymbol{x}_j) b_j(\boldsymbol{x}_j) \quad (2)$$

$$b_j(\boldsymbol{y}_j|\boldsymbol{x}_j) = N(\boldsymbol{y}_t|\boldsymbol{A}_j \boldsymbol{x}_j + \boldsymbol{\mu}_{\boldsymbol{y}_j}, \boldsymbol{\Sigma}_{\boldsymbol{y}_j}) \quad (3)$$

$$b_j(\boldsymbol{x}_j) = N(\boldsymbol{x}_t|\boldsymbol{\mu}_{\boldsymbol{x}_j}, \boldsymbol{\Sigma}_{\boldsymbol{x}_j}). \quad (4)$$

Here $\boldsymbol{q}$ denotes the state sequence shared by all features; π_{q_0} and $a_{q_{t-1} q_t}$ represent initial state probability and state transition probability from state q_{t-1} to q_t respectively; $b_j()$ denotes state observation probability density function (pdf) for state j; $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ represent mean vector and covariance matrix; and $\boldsymbol{A}_j$ denotes the linear projection matrix of $\boldsymbol{x}_j$ for state j. This state-wise linear transform approximates the non-linear dependency between two kinds of features. The expectation-maximization (EM) algorithm can be used to estimate the model parameters λ ($\boldsymbol{\mu}_{\boldsymbol{x}_j}$,$\boldsymbol{\Sigma}_{\boldsymbol{x}_j}$,$\boldsymbol{\mu}_{\boldsymbol{y}_j}$,$\boldsymbol{\Sigma}_{\boldsymbol{y}_j}$ and $\boldsymbol{A}_j$), for which details are given in [2].

The training procedures in [2] are as follows. Gaussian parameters ($\boldsymbol{\mu}_{\boldsymbol{x}_j}$,$\boldsymbol{\Sigma}_{\boldsymbol{x}_j}$,$\boldsymbol{\mu}_{\boldsymbol{y}_j}$,$\boldsymbol{\Sigma}_{\boldsymbol{y}_j}$) are initialized by separately learning two streams for spectrum and formant features with a shared decision tree. The linear transforms are then estimated as

$$\widehat{\boldsymbol{A}} = [\sum_{t=1}^{T} \gamma_j(t)(\boldsymbol{y}_t - \boldsymbol{\mu}_{\boldsymbol{y}_j})\boldsymbol{x}_t^\intercal)] \cdot [\sum_{t=1}^{T} \gamma_j(t)\boldsymbol{x}_t\boldsymbol{x}_t^\intercal)]^{-1} \quad (5)$$

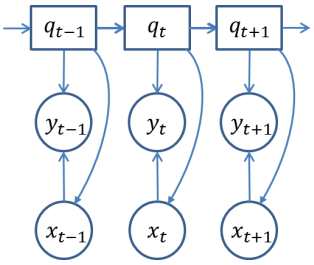

Figure 2: Modelling the dependency between formant ($\boldsymbol{x}$) and spectrum ($\boldsymbol{y}$) features. q denotes state.

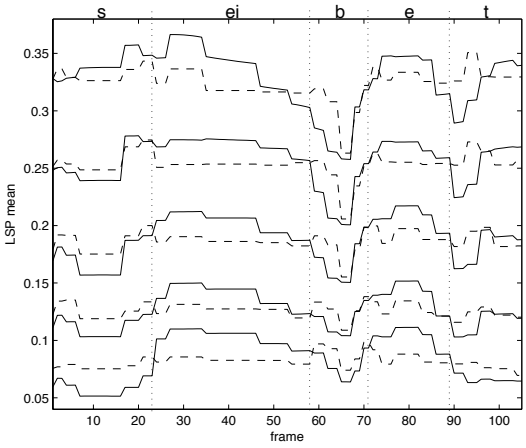

Figure 3: Comparison of trajectories (first 5 dimensions) of acoustic mean between $\boldsymbol{\mu}_{\boldsymbol{y}_j}$(dotted) and $\boldsymbol{A}_j\boldsymbol{x}_j + \boldsymbol{\mu}_{\boldsymbol{y}_j}$(solid).

where $\gamma_j(t)$ is the occupancy probability of state j at time t. We then re-estimate the Gaussian parameters based on the linear transforms estimated and repeat this process until convergence is reached.

However, this means the linear transforms are initialized as zero matrices (i.e. no dependency) which seems inappropriate. Even after several iterations, we have found the linear transforms may be close to the zero matrices. Therefore, in order to achieve better initialisation, here we use the following equation for the linear transforms and to start the estimation of the Gaussian parameters:

$$\widehat{\boldsymbol{A}} = [\sum_{t=1}^{T} \gamma_j(t)(\boldsymbol{y}_t - \boldsymbol{\mu}_{\boldsymbol{y}}^{\text{global}})\boldsymbol{x}_t^\intercal)] \cdot [\sum_{t=1}^{T} \gamma_j(t)\boldsymbol{x}_t\boldsymbol{x}_t^\intercal)]^{-1} \quad (6)$$

where $\boldsymbol{\mu}_{\boldsymbol{y}}^{\text{global}}$ is the global mean of spectrum feature over training data. Fig. 3 shows the comparison of $\boldsymbol{\mu}_{\boldsymbol{y}_j}$ and $\boldsymbol{A}_j\boldsymbol{x}_j + \boldsymbol{\mu}_{\boldsymbol{y}_j}$ in eq. (3) using the new initialisation, followed by a few iterations. Here LSPs are adopted as spectral features. From this figure we can see that $\boldsymbol{A}_j\boldsymbol{x}_j$ has a reasonable impact, especially in vowel regions. And also, the effect of using state-level linear transform to represent relations between LSPs and formants is shown in Fig. 3.

To manipulate the formants and affect the acoustic features, the maximum likelihood parameter generation algorithm given by

$$(\boldsymbol{X}_s^*, \boldsymbol{Y}_s^*) \approx \arg\max_{\boldsymbol{X}_s, \boldsymbol{Y}_s} P(\boldsymbol{W_X}\boldsymbol{X}_s, \boldsymbol{W_Y}\boldsymbol{Y}_s|\lambda, \boldsymbol{q}^*) \quad (7)$$

is approximated as two steps:

$$\boldsymbol{X}_s^* \approx \arg\max_{\boldsymbol{X}_s} P(\boldsymbol{W}_{\boldsymbol{X}} \boldsymbol{X}_s | \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}, \boldsymbol{q}^*) \quad (8)$$

$$\boldsymbol{Y}_s^* \approx \arg\max_{\boldsymbol{Y}_s} P(\boldsymbol{W}_{\boldsymbol{Y}} \boldsymbol{Y}_s | \boldsymbol{\mu}_{\boldsymbol{y}}, \boldsymbol{\Sigma}_{\boldsymbol{y}}, \boldsymbol{A}, f(\boldsymbol{X}_s^*), \boldsymbol{q}^*). \quad (9)$$

We generate formant trajectories in the sense of maximum likelihood first, and then we manipulate those formant trajectories. Here $f(*)$ denotes a function for manipulating the formant features. Finally, we generate spectral synthesis parameter trajectories in the sense of the maximum likelihood using the given manipulated formant trajectories.

3. Experiments

3.1. Experimental conditions

A male English speech database was used in our experiments, including 1,200 sentences for training and 63 sentences for test. These speech waveforms were recorded at 16kHz sample rate. The conventional acoustic features used for model training included F0 and spectral parameters, which were 40-order frequency-warped LSPs and an extra gain dimension derived from the spectral envelope obtained by STRAIGHT [7] analysis. The frame shift was set to 5ms. A 5-state left-to-right HMM structure with no skips was adopted to train context-dependent phone models, the minimum description length (MDL) criterion was adopted for decision tree building and the MDL factor was set to 1.0. As in [2], we have used the weighted RMSE of LSPs as an objective error measure. In addition, we have conducted subjective listening tests with 20 listeners. An acoustic-only system was built as a baseline, denoted as *Acou-only*.

Formant features were extracted using the Snack Sound Toolkit (http://www.speech.kth.se/snack/) with 5ms as frame length, to match the acoustic frameshift. These features comprised centre frequencies and bandwidths. The algorithm for formant extraction used in Snack applies dynamic programming to select and optimize a formant trajectory from multiple candidates which are obtained by solving for the roots of the linear predictor polynomial (poles of a synthesis filter). This means that, though the concept of a formant is mainly related to vowels, the formant features extracted by Snack for consonants too actually represent the roots of the linear predictor polynomial. Therefore, features for both vowels and consonants compose continuous formant trajectories. Currently, we only consider manipulation of $F1$ and $F2$, and so only $F1$ and $F2$ have been integrated into our system. $F1$ is frequently plotted against $F2-F1$ [9]. By plotting mean frequencies of vowels, equal distances along either axis are observed to correspond more closely to equal perceptual distance. Since both $F1$ and $F2$ are in the frequency domain, we have further applied the logarithm transform to them and adopted $\log F1$ and $\log(F2 - F1)$ as the static formant feature vector (i.e. $\boldsymbol{X}_s$). 100 transforms were adopted in our formant-controlled system, which is denoted as *Acou+Frm*.

3.2. Comparison with acoustic-only system

Although our goal is control over formants in synthesized speech, it is necessary also to compare the performance of the *Acou+Frm* system to that of the *Acou-only* system in terms of naturalness. In *Acou-only*, a total of 2,222 leaf nodes were split in MDL-based clustering, and 2,420 leaf nodes for *Acou+Frm*. This means integrating $F1$ and $F2$ does not introduce much distinction, in contrast to what happens when articulatory features are introduced [2]. This could be because the formant features have a close relationship to the spectral features, especially for LSPs. Table 1 shows RMSE of LSPs for the two systems. These two systems achieved very similar performance in terms of the objective measurement. Fig. 4 shows the preference score between two systems using 15 synthesized sentence pairs. We can see that there is no significant difference between the two. Since these two systems have similar numbers of leaf nodes, they should have similar performance. We conclude that *Acou+Frm* has similar performance to the conventional system (i.e. *Acou-only*) in terms of preference scores and objective measures.

Table 1: Comparison of RMSE of LSPs for *Acou-only* and *Acou+Frm*

RMSE of LSP	Training set	Test set
Acou-only	0.595	0.606
Acou+Frm	0.593	0.607

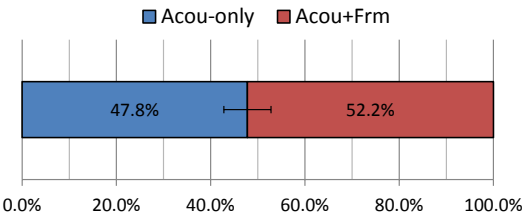

Figure 4: Preference score between *Acou-only* and *Acou+Frm*.

3.3. Evaluation of controllability

To evaluate controllability of *Acou+Frm* in a similar way to [2], we have chosen three front vowels /ɪ/, /ɛ/ and /æ/ in English for this experiment. As shown in Fig.1, the main difference among these three vowels in terms of formants is the position of $F1$ and the distance between $F2$ and $F1$. /ɪ/ has smallest $F1$ and largest $F2 - F1$, and /æ/ has largest $F1$ and smallest $F2 - F1$, and /ɛ/ has a mid position of $F1$ and middle distance of $F2 - F1$. Five monosyllabic words (bet, hem, led, peck, and set) with vowel /ɛ/ were selected and embedded into the carrier sentence "Please say ... again." According to the difference between /ɪ/ to /ɛ/ and /ɛ/ to /æ/ in the $F1$-$F2$ space, we adopted manipulating functions as shown in Table 2 to implement $f(*)$ in eq. (9):

Table 2: Formant manipulating functions $f(*)$

label	-3	-2	-1	0	+1	+2	+3
$F1$ (Hz)	+150	+100	+50	0	-100	-200	-300
$F2$ (Hz)	-300	-200	-100	0	+100	+200	+300

Each monosyllabic word was manipulated to create a total of 7 degrees of modification. Twenty listeners were then asked to listen to the synthesized sentences (35 sentences in total) and write down the key word in the carrier sentence. Then, for each manipulation degree, we calculated the percentage of how these three vowels were perceived.

Fig. 5 shows perception of vowels before and after manipulating formants in the proposed model[1]. It is very clear that /ɛ/ is changed to /ɪ/ or /æ/ if $F1$ and $F2$ are manipulated appropriately, e.g. method +2, +3, -2 and -3. This shows it is possible

[1]Some samples can be found in `http://home.ustc.edu.cn/~leiming/Demo_IS2011/Demo.html` or `http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/demo.html`

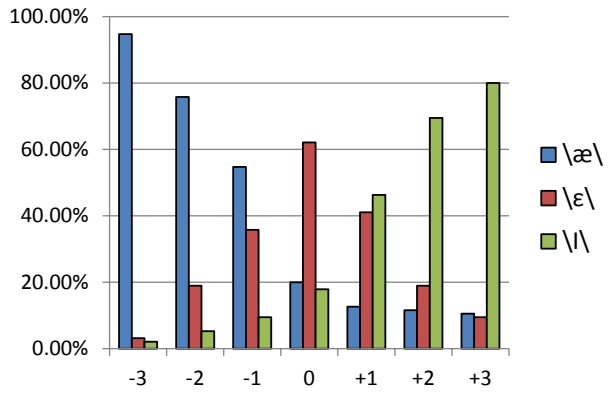

Figure 5: *Vowel perception for each manipulation method.*

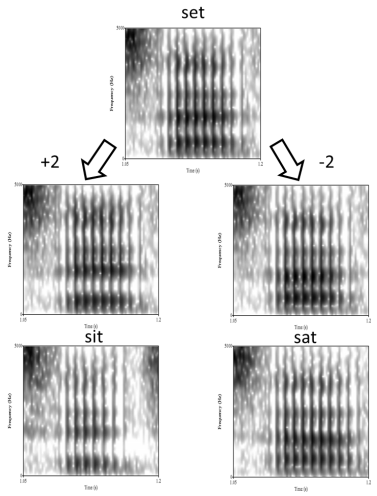

Figure 6: *Effect of manipulation method -2 and +2 on spectrogram of "set".*

to control vowel pronunciation in the speech synthesized by our system according to phonetic knowledge of formants. Fig. 6 shows the effect of manipulation degrees -2 and +2 on the spectrogram of /ɛ/ in "set". The effect on $F2 - F1$ is discernable: it become smaller and larger in degrees -2 and +2 respectively. After manipulation, the spectrogram of "set" is close to that of either "sat" or "sit".

Finally, to verify the quality of the manipulated vowels, we selected manipulation methods -2 and +2 (which are phonetically correct for the vowel differences) as new target vowels of /æ/ and /ɪ/, to compare with the corresponding sentences generated by the *Acou-only* system. Each monosyllabic word was manipulated to create 2 variants, and in total 10 sentence pairs were used as part of the subjective listening test. Fig. 7 shows the preference score between *Acou-only* and the manipulated *Acou+Frm* system. Comparing this with Fig. 4, we can see that there is no significant difference between the two.

4. Conclusion and Future work

This paper has proposed a novel framework that enables us to manipulate and control formants in HMM-based speech synthesis. The advantages of the proposed method are that 1) formants affect only statistical models trained on spectral envelop features (such as LSPs), and 2) speech is generated not by formant

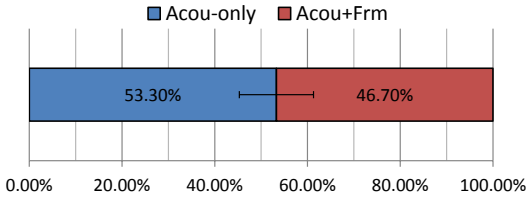

Figure 7: *Preference score between* Acou-only *and manipulated* Acou+Frm*.*

synthesisers but by high-quality vocoders using the spectral envelope features generated from statistical models with the influence of formants. The first advantage allows us to control the generation of acoustic features by manipulating a formant trajectory. We can achieve spectral modification by changing only those parts where we want control, whereas the user must specify all formants manually in conventional formant-based synthesisers. The second advantage provides us with high-quality speech. Taken together, this synthesiser will be very useful in other fields, such as speech perception and phonetics research, where the the Klatt model is currently the main tool. The experimental results have also shown the proposed method offers control over vowels in synthesized speech via the manipulation of $F1$ and $F2$, without any degradation of synthesis quality.

Integrating richer formant features, e.g. $F3$, $F4$, $F5$ and their bandwidths, as well as an investigation into different numbers of linear transforms, will be the subject of future work.

Acknowledgements The research leading to these results was partly funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 256230 (LISTA), and by EPSRC grant EP/I027696/1. This work was partially funded by the National Nature Science Foundation of China (Grant No. 60905010).

5. References

[1] K. Tokuda, H. Zen, and A. W. Black, "HMM-based approach to multilingual speech synthesis," in *Text to speech synthesis: New paradigms and advances*, S. Narayanan and A. Alwan, Eds. Prentice Hall, 2004.

[2] Zhen-Hua Ling, Korin Richmond, Junichi Yamagishi, and Ren-Hua Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *Trans. Audio, Speech and Lang. Proc.*, vol. 17, pp. 1171–1185, 2009.

[3] D.H Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, pp. 971–995, 1980.

[4] James L. Flanagan, Jont B. Allen, and Mark A. Hasegawa-Johnson, *Speech Analysis Synthesis and Perception*, Springer-Verlag, 2008.

[5] Hongwei Hu and Martin J. Russell, "Improved modelling of speech dynamics using non-linear formant trajectories for HMM-Based speech synthesis," in *Proc. of InterSpeech*, 2010.

[6] A. Acero, "Formant analysis and synthesis using hidden Markov models," in *Proc. of EUROSPEECH*, 1999.

[7] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using pitch-adaptive time-frequency smoothing and an instanta-neous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[8] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. of ICASSP*, 1995, vol. 1, pp. 660–663.

[9] Peter Ladefoged and Ian Maddieson, *The Sounds of the World's Languages*, Wiley-Blackwell, 1996.