Formant Controllable HMM-based Speech Synthesis

Ming Lei*[1,3],Catherine Mayo[2],Korin Richmond[2], Junichi Yamagishi[2], Zhen-Hua Ling[1], Simon King[2]

[1]*National Engineering Laboratory of Speech and Language Information Processing, University of Science and Technology of China, Hefei, China*

[2]*Center for Speech Technology Research, University of Edinburgh, Edinburgh, UK*

[3]*Search Technology Center Asia, Microsoft Corp., Beijing, China*

Abstract

Hidden Markov model (HMM) based statistical parametric speech synthesis (SPSS) has become a very popular method for text-to-speech conversion. The method gives good performance in terms of naturalness, and reproduces many of the acoustic characteristics of the human voice. Unfortunately, the method is also rather opaque, in that the acoustic relevance, or "meaning", of each of the huge number of model parameters is far from obvious. Recent work has shown it is possible to integrate articulatory features within an HMM-based SPSS system, and to control the synthetic speech in terms of those articulatory features. Articulatory measurement data reflects the physical speech production mechanism, which offers an easily-understandable domain for controlling synthetic speech, but which is also relatively inconvenient to acquire. Formant features have a straightforward relationship to vocal tract configurations but, unlike articulatory data, can be easily obtained by estimating directly from recorded speech waveforms. The purpose of this paper is to investigate the integration of formant features into state-of-the-art HMM-based SPSS. By modeling the relationship between formant features and the spectral features of the synthesis vocoder, we aim to introduce control over the synthesized speech via formant features that are predicted in parallel during the synthesis process. We have conducted two categorical perception experiments to evaluate and analyse the controlability of this approach quantitatively. The results we present show that the synthetic speech may be very readily controlled in terms of simple formant features, and that prior phonetic knowledge of formants may be easily applied.

Keywords: speech synthesis, hidden Markov model, formant features, formant synthesizer, controlability, speech perception

1. Introduction

Over the past five years, hidden Markov model (HMM) based statistical parametric speech synthesis (SPSS) has become a mainstream approach to text-to-speech conversion, offering a high degree of flexibility and naturalness (Zen et al., 2009). HMM-based SPSS is an automatic, data-driven method for producing smooth and stable synthetic speech. Its use of a powerful statistical model brings several advantages. The inherent flexibility of the model means that very little explicit knowledge of speech production is needed, as it can capture the acoustic characteristics of speech such as emphasis, varying speaking styles, personality and emotion automatically. It opens up numerous possibilities in the application of speech synthesis technology, for example transforming voice characteristics, speaking styles and emotions (Yamagishi and Kobayashi, 2007; Shichiri et al., 2002; Tachibana et al., 2005). It can accommodate a wide variety of languages, and also scales trivially to use large amounts of data (King and Karaiskos, 2010).

Despite this impressive flexibility, however, current HMM-based SPSS methods do not allow structured prior knowledge of speech production and perception to be incorporated in a straightforward way. In essence, this is because the acoustic signal is modeled *implicitly*. Though many aspects of human speech have been subjected to experimental investigation, the knowledge that has accrued cannot be directly applied within the current framework in order to control synthetic speech at will, since the prior knowledge that is related to the spectral features of state-of-the-art vocoders is limited. For example, studies have indicated that to improve the intelligibility of synthetic speech, one can reduce formant bandwidths or mimic formant effects observed in hyperarticulated speech (Picart et al., 2011), or increase vertical displacement of articulator movements (Ling et al., 2009). But it is far from clear how to manipulate the acoustic characteristics of synthesised speech by changing the 24- or 40-order spectral features found in a typical HMM-based SPSS system, for example. All physiological and perceptual "meaning" is obscured in this parameterization. Although a few simple feature enhancement methods (Ling et al., 2006) have been reported, it remains very difficult to integrate arbitrary prior knowledge to achieve more complex manipulations.

Recently, a set of work which aims to integrate articulatory features into HMM-based SPSS has been reported. This work

*M. Lei is now with Microsoft Corp., Beijing, China. Most of this work was carried out when he was a PhD candidate at University of Science and Technology of China.

Pilot work for this research was presented at Interspeech (Lei et al., 2011).

has shown that it is in fact possible to exploit explicit knowledge of speech production to manipulate synthesized speech (Ling et al., 2009, 2011). Under this approach, the conventional spectral features of the vocoder are modeled jointly with articulatory features, including tongue movements captured using electromagnetic articulography (EMA), within the HMMs. The relationship between these conventional spectral features and the articulatory features is modelled by linear transforms. To perform synthesis, the articulatory feature sequences are predicted first, and then these may be manipulated at will according to prior knowledge of articulator movements. The vocoder spectral features are then predicted under the influence of these manipulated articulatory features, according to the learned piecewise-linear relationship, and finally the acoustic speech signal may be obtained from these. This approach makes it possible to utilize prior knowledge pertaining to articulatory features, and for those modifications to be reflected in the subsequently synthesized acoustic signal.

Although the work presented in Ling et al. (2009, 2011) has shown it is possible to impose more transparent and meaningful control over an HMM-based SPSS system, the use of articulatory features is not ideal for all cases, since they can be relatively difficult to acquire. Specialist equipment is required to record the movements of articulators together with the acoustic speech signal. Articulography techniques may impact speech production in certain ways, and in some cases it may not be possible to record articulatory data at all (e.g. with child subjects). With these drawbacks in mind, we propose it would be useful to investigate the possibility of using a similar approach, but with a different set of features. Ideally, these should i) be convenient to acquire; ii) provide a good representation of the vocal tract; and iii) have a strong foundation of prior knowledge about them.

With this in mind, we note formant features can also characterize many aspects of speech in a compact and meaningful way, but have the advantage, in contrast to articulatory features, that they can be estimated directly from the speech waveform. As acoustic resonances of the human vocal tract, formant features have several attractive properties. For example, $F1$ (Fj denotes the jth formant central frequency) is related to vowel openness, and $F2$ corresponds to frontness, which are crucial characteristics of vowels. Vowels are distinguishable, and perceived as different, due to differing formant center frequencies (Flanagan et al., 2008), while the transitions of formant center frequencies in a vowel onset can influence the perception of a preceding stop (Delattre et al., 1955). Speech variability arising from differences in gender, dialect, age, emotion, and degree of hypo- or hyper-articulation can be investigated directly by transform of formant feature space (Jacewicz et al., 2007; Hawkins and Midgley, 2005; Kienast and Sendlmeier, 2000; Picart et al., 2011). Variations typically observed when a speaker is talking in a noisy environment compared to a quiet one, referred to as the Lombard effect, can be described in terms of formant feature changes (Lane and Tranel, 1971). Formant features, often in conjunction with the classic formant synthesizer, have been used extensively in the field of speech perception research (Ladefoged and Broadbent, 1957; Kewley-Port and Zheng, 1999; Liu and Kewley-Port, 2004; Kawahara et al., 2008; Klatt, 1980), which further underlines the usefulness and efficacy of formant features. These examples suggest formant features offer a promising replacement for articulatory features as a suitable domain in which to exert control over the generation of speech in an HMM-based SPSS system. Moreover, it is possible a formant-controlled HMM-based speech synthesizer could prove beneficial for speech perception research, since an HMM-based SPSS has many advantages compared to the use of a classical formant synthesizer.

Some previous work has been reported on the use of formant features in an HMM-based SPSS system (Hu and Russell, 2010; Acero, 1999). In such work, formant features are directly used to characterize the speech spectrum, and formant features are predicted automatically by the trained model as input for a conventional formant synthesizer to produce speech. However, reliance upon a formant synthesizer does not allow fine spectral detail to be modeled to the same degree as with a general HMM-based SPSS system, and so synthetic speech cannot be produced with the same level of naturalness.

Previously, we have presented preliminary work on integrating formant features into HMM-based SPSS (Lei et al., 2011), in which $F1$ and $F2$ were selected to represent the vocal tract configuration. Using a method similar to Ling et al. (2009), we showed the relationship between formant features and the conventional spectral features used for the resynthesis vocoder can be modeled, and that by controling the formant features the synthesized speech can be modified in ways that match our expectations according to prior knowledge of formants. In this paper, we expand upon that preliminary work by presenting a fuller and more rigorous investigation. There are several key differences. First, richer formant features are used here[1]. Second, we evaluate a newer method (compared to that used in Lei et al., 2011) to jointly model the formant and vocoder spectral features. This method, adopted from Ling et al. (2011), involves the introduction of an additional Gaussian mixture model to tie the linear transforms that relate the formant features to the conventional spectral features used for the resynthesis vocoder. We compare the performance of this new method with the previous method of context-dependent state-based tying implemented using a clustering decision tree. Finally, we present results of perceptual experiments that have been designed as a more extensive and rigorous test of the proposed methods, similar to those typically conducted in the field of speech perception research. Overall, the purpose of this paper is to introduce and rigorously evaluate an HMM-based SPSS system which may be controled in terms of formants, which is more convenient to construct than the articulatory-controlable HMM-based SPSS system presented in Ling et al. (2011), and thus which offers a new tool for speech perception research as an alternative to the formant synthesizer.

[1] Specifically, we refer to formant center frequencies. Formant bandwidth and amplitude will be ignored, since formant center frequency is the most salient feature for our purpose.

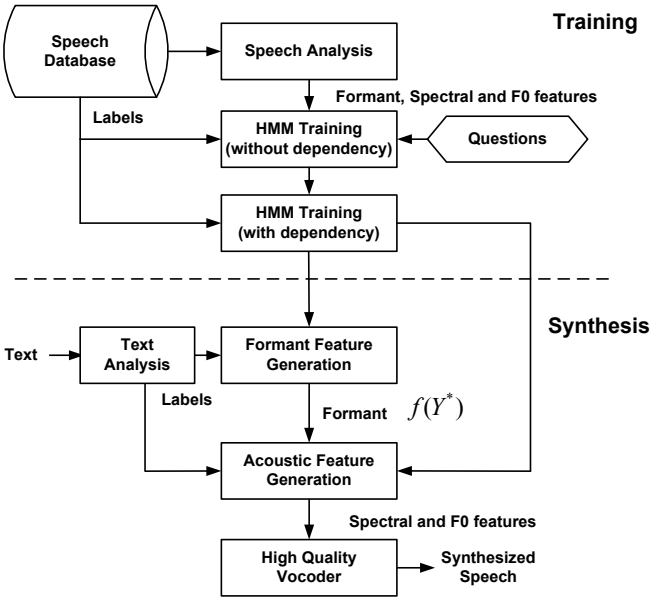

Figure 1: The proposed formant-controllable HMM-based SPSS system.

2. Formant Controllable HMM-based Speech Synthesis

Under the approach previously proposed in Ling et al. (2009, 2011), articulatory features are integrated into HMM-based SPSS by training a unified HMM model with parallel acoustic and articulatory observations. The dependency between the acoustic and articulatory streams is represented by a set of linear transforms, which may either be tied in terms of context-dependent states (Ling et al., 2009) or tied in the articulatory feature space (Ling et al., 2011). To perform synthesis, the articulatory features are generated from the unified HMM first, and these may then be manipulated arbitrarily in accordance with specific phonetic knowledge. The acoustic features are then predicted under the influence of the manipulated articulatory features.

In this paper, the same framework of cross-stream dependency modelling is adopted to achieve a controllable HMM-based SPSS driven by formant features. The model training and parameter generation algorithms will be briefly reviewed in this section.

2.1. Model Training

An overview of the proposed formant-controllable HMM-based SPSS system is shown in Fig. 1. During training, an HMM λ for the combined spectral and formant features is estimated under the maximum likelihood criterion. Here, the term "spectral features" refers to the acoustic features used in a standard HMM-based SPSS system to represent the spectral envelope of a speech frame, such as mel-cepstra or line spectral pairs (LSP), from which a speech waveform may be synthesised. The formant features could comprise any of the formant center frequencies or bandwidths at each frame, for example. Let $\boldsymbol{X} = [\boldsymbol{x}_1^\intercal, \boldsymbol{x}_2^\intercal, \ldots, \boldsymbol{x}_T^\intercal]$ and $\boldsymbol{Y} = [\boldsymbol{y}_1^\intercal, \boldsymbol{y}_2^\intercal, \ldots, \boldsymbol{y}_T^\intercal]$ denote the spectral feature sequence and the formant feature sequence of T frames respectively. For each frame, $\boldsymbol{x}_t \in \mathcal{R}^{3D_X}$ and $\boldsymbol{y}_t \in \mathcal{R}^{3D_Y}$ are composed of static, delta and acceleration components (Tokuda et al., 2000), where D_X and D_Y are the dimensionality of the static spectral features and formant features respectively. Therefore, we have $\boldsymbol{X} = \boldsymbol{W_X}\boldsymbol{X}_s$ and $\boldsymbol{Y} = \boldsymbol{W_Y}\boldsymbol{Y}_s$, where $\boldsymbol{W_X}$ and $\boldsymbol{W_Y}$ are the matrices used to calculate observations from static feature sequences $\boldsymbol{X}_s$ and $\boldsymbol{Y}_s$ (Tokuda et al., 2000). The unified HMM λ is estimated by maximising

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

where π_j and a_{ij} represent initial state probability and state transition probability; $b_j(\cdot)$ is the state observation probability density function (PDF) for state j; and $\boldsymbol{q} = \{q_1, q_2, ..., q_T\}$ denotes the state sequence shared by both spectral and formant features.

Assuming formants to be fundamental acoustic features intrinsic to the characteristics of the vocal tract filter, $b_j(\boldsymbol{x}_t, \boldsymbol{y}_t)$ may be decomposed as

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

where

$$b_j(\boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{\boldsymbol{y}_t}, \boldsymbol{\Sigma}_{\boldsymbol{y}_t}), \quad (3)$$

and $\mathcal{N}(;\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. Because the relationship between formant features and spectral envelope features is non-linear and complex, a set of multiple linear transforms, which constitute a piecewise linear transform, is introduced to model $b_j(\boldsymbol{x}_t|\boldsymbol{y}_j)$. Furthermore, the decision as to which of the set of linear transforms is "active" for any given frame in an utterance is governed by a tying strategy. Two different strategies have been proposed to mediate this cross-stream dependency:

1) *Context-dependent transform tying*. In this model structure, after Ling et al. (2009), different transform matrices are applied to different HMM states, as shown in Fig. 2b). Because the state PDFs are trained context-dependently in our method, as is standard for HMM-based SPSS, the linear transforms are also context-dependent. Let $\boldsymbol{A}_j$ denote the linear transform matrix for state j. Then, $b_j(\boldsymbol{x}_j|\boldsymbol{y}_j)$ can be written as

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

By introducing the new term $\boldsymbol{A}_j\boldsymbol{y}_t$ which can influence the statistical mean of the spectral features at each frame, the formant feature vector $\boldsymbol{y}_t$ can affect the distribution of the spectral features. An expectation-maximization (EM) based model training algorithm has been developed to estimate all model parameters $\{\boldsymbol{\mu}_{\boldsymbol{x}_j}, \boldsymbol{\Sigma}_{\boldsymbol{x}_j}, \boldsymbol{\mu}_{\boldsymbol{y}_j}, \boldsymbol{\Sigma}_{\boldsymbol{y}_j}, \boldsymbol{A}_j\}$ (Ling et al., 2009). In order to reduce the number of parameters that need to be estimated, the state-dependent transform matrices $\boldsymbol{A}_j$ are further tied to a given class using the decision tree for spectral model clustering. Each matrix is defined as a three-block form corresponding to static, velocity and acceleration components.

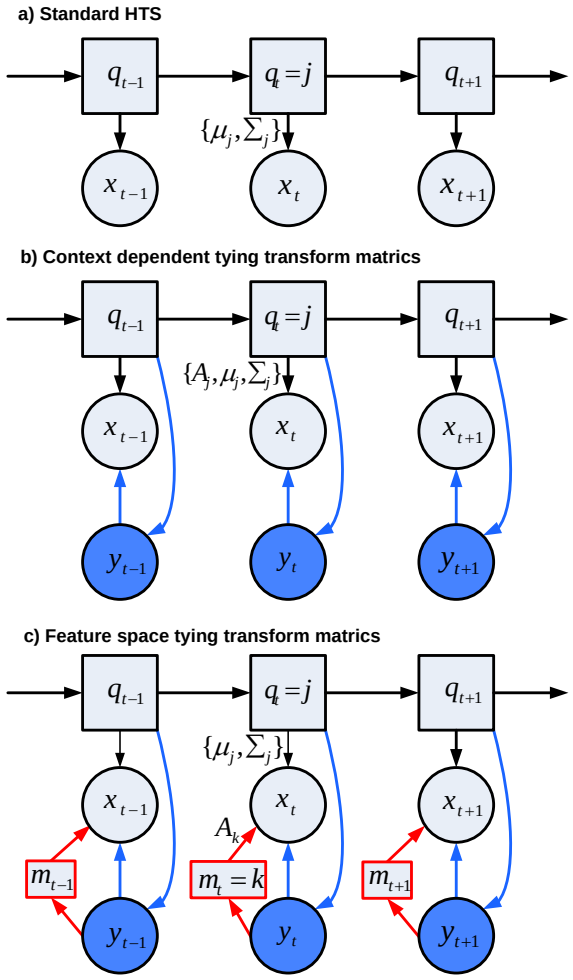

Figure 2: Model structures for a) standard HMM-based SPSS, b) formant-controllable HMM-based SPSS using *context-dependent* transform tying, and c) formant-controllable HMM-based SPSS using *feature-space* transform tying. Vectors $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ denote the spectral and formant features at the t-th frame respectively.

2) *Feature-space transform tying*. In this model structure, after Ling et al. (2011), the transform matrix is determined by formant features instead of the context information, as shown in Fig. 2c). First, a Gaussion mixture model (GMM) $\lambda^{(G)}$ containing M mixture components is fitted to the formant features of the training data. Then, a transform matrix is estimated for every mixture component of $\lambda^{(G)}$. Here, $b_j(\boldsymbol{x}_j|\boldsymbol{y}_j)$ is defined as

$$b_j(\boldsymbol{x}_j|\boldsymbol{y}_j) = \sum_{k=1}^{M} P(\boldsymbol{x}_t, m_t^{(G)} = k|\boldsymbol{y}_t, q_t = j, \lambda, \lambda^{(G)}) \quad (5)$$

$$= \sum_{k=1}^{M} \zeta_k(t) P(\boldsymbol{x}_t|\boldsymbol{y}_t, q_t = j, m_t^{(G)} = k, \lambda, \lambda^{(G)}), \quad (6)$$

where $m_t^{(G)}$ denotes the mixture index of $\lambda^{(G)}$ for formant feature vector at frame t;

$$\zeta_k(t) = P(m_t^{(G)} = k|\boldsymbol{y}_t, q_t = j, \lambda, \lambda^{(G)}) = P(m_t^{(G)} = k|\boldsymbol{y}_t, \lambda^{(G)}) \quad (7)$$

because the HMM state sequence $\boldsymbol{q}$ and the GMM mixture sequence $\boldsymbol{m}^{(G)} = \{m_1^{(G)}, m_2^{(G)}, \ldots, m_T^{(G)}\}$ are assumed to be independent. For each Gaussian mixture, the dependency between the two feature streams is represented as

$$P(\boldsymbol{x}_t|\boldsymbol{y}_t, q_t = j, m_t^{(G)} = k, \lambda, \lambda^{(G)}) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{A}_k\boldsymbol{\xi}_t + \boldsymbol{\mu}_{\boldsymbol{X}_j}, \boldsymbol{\Sigma}_{\boldsymbol{X}_j}) \quad (8)$$

where $\boldsymbol{\xi}_t = [\boldsymbol{y}_t^\intercal, 1]^\intercal \in \mathcal{R}^{3D_Y+1}$ is the expanded formant feature vector and $\boldsymbol{A}_k \in \mathcal{R}^{3D_X \times (3D_Y+1)}$ is the transform matrix for the k-th mixture of $\lambda^{(G)}$. All parameters are estimated by the EM algorithm, the details of which are to be found in Ling et al. (2011).

As indicated in Fig. 1, the model training procedure comprises the following two main steps.

1) *Model training without cross-stream dependency*. Conventional two-stream HMM training is first applied to initialize $\{\boldsymbol{\mu}_{\boldsymbol{x}_j}, \boldsymbol{\Sigma}_{\boldsymbol{x}_j}, \boldsymbol{\mu}_{\boldsymbol{y}_j}, \boldsymbol{\Sigma}_{\boldsymbol{y}_j}\}$ for subsequent training which takes into account cross-dependency modelling. Shared decision trees are built to tie model parameters of both spectral and formant features under the minimum description length (MDL) criterion (Shinoda and Watanabe, 2000).
2) *Model training with cross-stream dependency*. The dependency between the spectral and formant features is represented by a set of linear transforms, which are estimated either by context-dependent tying or by feature-space tying, as discussed above. The initial $\boldsymbol{A}_j$ in (4) or $\boldsymbol{A}_k$ in (8) is estimated by fixing $\boldsymbol{\mu}_{\boldsymbol{X}_j} = 0$ for each HMM state in the EM re-estimation process.

Fig. 3 shows the static component of the acoustic mean vector $\boldsymbol{\mu}_{\boldsymbol{X}_j}$ in a middle state j of one HMM (for the /æ/ vowel) before and after estimating transform matrix $\boldsymbol{A}_j$. We can see that after we estimate $\boldsymbol{A}_j$, the static part of the acoustic mean $\boldsymbol{\mu}_{\boldsymbol{x}_j}$ in (4) is very close to 0 for all LSP coefficients, meaning that most acoustic information is captured by $\boldsymbol{A}_j\boldsymbol{y}_t$, and hence it is expected that manipulation of the formant features can exert a significant impact on the spectral features.

2.2. Parameter Generation with Formant Control

In order to achieve direct control over the formant characteristics of synthetic speech, the maximum likelihood parameter generation algorithm given optimal state sequence $\boldsymbol{q}^*$ is approximated as two steps (Ling et al., 2009):

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

$$\boldsymbol{X}_s^* \approx \arg\max_{\boldsymbol{X}_s} P(\boldsymbol{W}_X\boldsymbol{X}_s|\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}, \boldsymbol{A}, f(\boldsymbol{Y}_s^*), \boldsymbol{q}^*). \quad (10)$$

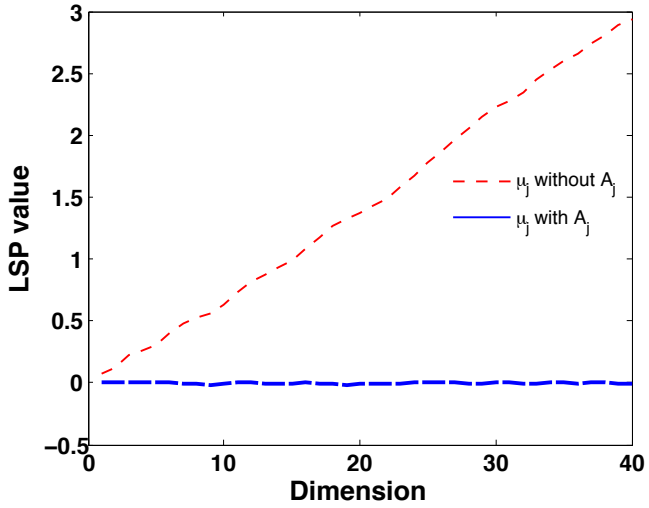

Figure 3: The acoustic mean vector of the static component of the spectral features in the middle state of one HMM (for vowel /æ/) without and with transform matrices $\boldsymbol{A}_j$ for the formant-controllable system using context-dependent transform tying.

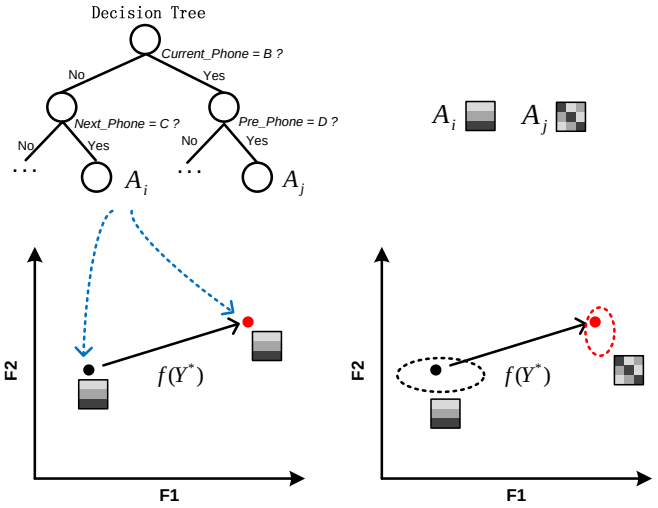

Figure 4: Comparison between the context-dependent (left) and feature-space (right) transform tying approaches to parameter generation with formant manipulation. The black arrow lines denote formant feature manipulation. The black and red points denote the original and manipulated features in the formant feature space respectively. The feature space tying approach allows the transform to change in response to formant manipulation, whereas context-depending tying does not enable this.

where $\boldsymbol{Y}_s^*$ and $\boldsymbol{X}_s^*$ denote the generated formant and spectral feature sequences respectively; and $f(\cdot)$ denotes a function for manipulating the generated formant trajectories according to specific phonetic knowledge. The complete parameter generation procedure may be summarized as follows:

1) Generate the optimal state sequence $\boldsymbol{q}^*$ using the trained duration model.
2) Generate the optimal formant features $\boldsymbol{Y}_s^*$ using only the formant stream in the trained HMM, so as to maximize (9).
3) Manipulate the generated formant features from $\boldsymbol{Y}_s^*$ to $f(\boldsymbol{Y}_s^*)$ according to the specific formant control task at hand.
4) Generate the optimal spectral feature sequence $\boldsymbol{X}_s^*$ under the influence of the manipulated formant features (10).

Detailed parameter generation formulae for both the context-dependent and feature-space transform tying approaches can be found in Ling et al. (2009, 2011). The difference between these two approaches in parameter generation with formant control is illustrated in Fig. 4. In the context-dependent transform tying approach, the transform matrix is entirely determined by the context description of each HMM state and does not change when the formant features are manipulated within the formant feature space. While in the feature-space transform tying approach, $\zeta_k(t)$ in (7) varies according to $\boldsymbol{y}_t$, which means the mixed linear transform in (6) can change adaptively to a new form that is more suitable for the manipulated formant features. The advantage of the feature-space transform tying approach over the context-dependent linear transforms when using articulatory control has been proved in our previous work (Ling et al., 2011). Here, we compare the effectiveness of these two approaches for the formant-controllable HMM-based SPSS system proposed and evaluated here.

3. System configurations

The proposed speech synthesis systems were trained on a speech database consisting of 1200 sentences available for training purposes , uttered by a male British English speaker. These speech waveforms were recorded at 16kHz sample rate with 16 bit precision in a quiet room with some sound treatment.

The conventional acoustic features included fundamental frequency (F0) and spectral parameters, which comprised 40-order LSPs and an extra gain dimension derived from the spectral envelope obtained by STRAIGHT analysis. The frame shift of the STRAIGHT analysis and F0 extraction was set to 5ms.

The formant features used in the proposed systems were extracted using the Snack Sound Toolkit (Sjolander et al., 1998) which includes a formant tracker based on LPC polynomial roots and dynamic programming. Note that the LPC order used for the formant tracker is significantly lower than that of the LPC/LSP analysis used as the above spectral parameters and therefore the relation between two features is nonlinear. The frame shift was set to be 5ms in order to keep formant features synchronized with the other acoustic features. Although the Snack Sound Toolkit can estimate formant center frequencies together with their respective bandwidths, we have only taken the formant center frequencies of $F1$, $F2$, and $F3$ into consideration. These are the most important acoustic cues in our experimental design mentioned later. We have opted to model these in the logarithmic frequency domain as the formant features; in other words the formant features we used are $\log F1$, $\log F2$, and $\log F3$.

A 5-state left-to-right multi-stream HMM structure with no skip paths was adopted to train context-dependent phoneme models. For each state and each feature stream, the minimum

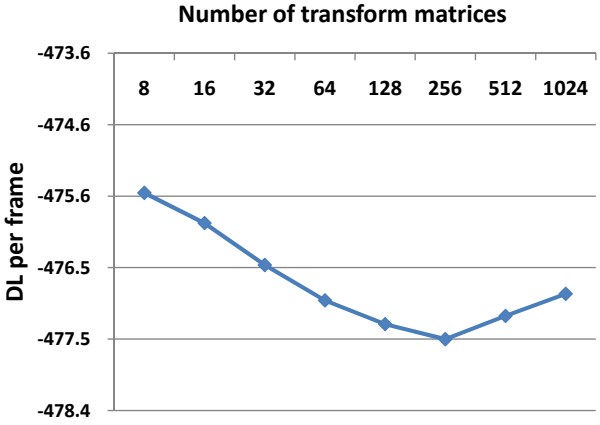

Figure 5: Description length per frame on the training set with varying numbers of transform matrices for the *Frm-GMM* system.

description length (MDL) criterion was applied to guide decision tree building. The proposed formant controllable systems were trained using LSPs, F0, formant features, and again their respective delta features. In order to impose the same decision tree structures, shared decision trees were built across the LSP and formant-feature streams. Here, we explicitly split the root node of the decision trees according to the question *"Is the current phone a vowel?"*. This was to guarantee the separation of vowel and consonant models, and to allow us to achieve vowel and consonant control completely separately. For the proposed formant-controllable systems, we built two systems to compare the transform tying methods. The systems we evaluated are denoted as follows:

- *Frm-DT*: the formant-controllable system in which the transform matrices are shared using the context-decision-tree based tying method.
- *Frm-GMM*: the formant-controllable system in which the transform matrices are shared using the GMM-based feature-space tying method.

The number of transform matrices used in the *Frm-GMM* system was adjusted based on the MDL criterion (Shinoda and Watanabe, 2000). Here, description length is defined as

$$DL(\lambda) \equiv -\log P(\boldsymbol{X}, \boldsymbol{Y}|\lambda) + \frac{1}{2} D(\lambda) \log G + C \qquad (11)$$

where $\log P(\boldsymbol{X}, \boldsymbol{Y}|\lambda)$ is the log likelihood function of the model for the training set; $D(\lambda)$ is the dimensionality of the model parameters; G is the total number of observed frames in the training set; and C is a constant. Considering the three-block matrix structure of $\boldsymbol{A}_j$, $D(\lambda) = 3M(D_{\boldsymbol{X}}-1)(D_{\boldsymbol{Y}}+1)+C_D$, where the power dimension is omitted from $\boldsymbol{A}_j$, and C_D is a constant that is independent from the number of transform matrices M.

We calculated description length per frame on the training set for various values of M. The results are shown in Fig. 5. We see from this figure that $M = 256$ leads to the minimum description length. Thus, we chose to use 256 transform matrices for the *Frm-GMM* system, and hence 256 mixture components were used to model the formant feature space. The number of transform matrices used for the *Frm-DT* system was set to 275. Specifically, tying transform matrices were applied at nodes to a depth of 7 (where the root node has a depth 0) within each state decision tree, in order to make this system comparable to the *Frm-GMM* system.

4. Formant control experiments

The proposed method allows prior knowledge of formant behaviour to be used to manipulate the synthetic speech output by an HMM-based speech synthesis system. To assess the effectiveness of controlling the proposed systems using formant parameters in this way, we have designed two perception experiments involving vowel and consonant identification tasks respectively. For both these tasks, perception is predominantly affected by the first three formant center frequencies. [2]

4.1. Vowel contrast stimuli

For the vowel identity perception experiment, a three-way contrast of /bɪt/ ("bit"), /bɛt/ ("bet"), and /bæt/ ("bat") was created. This was achieved by manipulating the generated formant features of the synthetic word "bet" in the target vowel region /ɛ/, and then the conventional spectral features for resynthesis were predicted under the influence of these manipulated formant features.

Both sets of stimuli (from systems *DT-HMM* and *GMM-HMM*) varied along a 7-point continuum of $F1$, $F2$ and $F3$ vowel formant **steady-state** values. The continuum was created by synthesising a /bɛt/ token to serve as the midpoint of the continuum, and then manipulating the vowel portion of this token's formants in equal steps to reach /bɪt/ at one end of the continuum, and /bæt/ at the other end of the continuum. Specifically, from analysis of the speech data on which the systems were trained, an unmodified /bɛt/ had steady-state formant values around F1=530Hz, F2=1840Hz, F3=2550Hz. To create the continuum towards /bɪt/, $F1$ and $F3$ were adjusted down in 3 steps (50Hz for each step) and $F2$ was adjusted up in 3 steps (50Hz for each step), resulting in intended formant values for /bɪt/ of $F1 = 380Hz$, $F2 = 1990Hz$ and $F3 = 2400Hz$. We have confirmed that these are consistent with natural formant values for /bɪt/. Similarly, to create the continuum towards /bæt/, the steady-state formant values of the original /bɛt/ were adjusted as follows: $F1$ and $F3$ we adjusted up in 3 steps, while $F2$ was adjusted down in 3 steps (50Hz for each step). This resulted in intended formant values for /bæt/ of $F1 = 680Hz$, $F2 = 1690Hz$, $F3 = 2400Hz$. Again, these values are consistent with natural formant values for /bæt/.

4.2. Consonant contrast stimuli

A three-way consonant contrast /bɛt/ ("bet"), /dɛt/ ("det") and /gɛt/ ("get") was selected for our consonant perception experiments. The consonant contrast stimuli were designed to

[2] Some examples of the synthetic speech used in the experiments can be found at `http://staff.ustc.edu.cn/~zhling/FrmCtrl/demo.html`.

vary along a 9-point continuum of $F1$, $F2$ and $F3$ vowel formant **onset** values. This continuum was created by first synthesising a /dɛt/ token to serve as the midpoint of the continuum. Then, since it is the formant trajectories, and not the **steady-state** values, that will influence consonant perception, we used interpolation to obtain the target vowel formant trajectories. To achieve this, we first synthesized "bet", "det" and "get" without any manipulation, and all with a duration to match that of the "det" token, to give synchronized target formant trajectories. Then, to create the endpoint /bɛt/ token, the /dɛt/ token was resynthesised using the vowel formant contour, including onset part, found in the unmodified /bɛt/ token. Similarly, the endpoint /gɛt/ token was created by resynthesising the /dɛt/ token with the vowel formant contour from /gɛt/. To create intermediate points on the continuum between each of these canonical tokens, the vowel formant contours were linearly interpolated in 5 equal-sized steps between those found in /dɛt/ and those found in /bɛt/, and again between those found in /dɛt/ and those found in /gɛt/. The interpolation and resynthesis was done in order to maintain consistency of all of the acoustic-phonetic details of the tokens, except onset transitions, across the continuuum.

Though these three consonants are distinguishable in terms of their typical formant trajectories (Delattre et al., 1955), it is well known that burst characteristics can also influence the perception of stop consonants in addition to formant transitions. Since our focus here is on listeners' perceptual processing of manipulated vowel formant values, we chose to use burst-less synthetic consonant stimuli in this study, as in Harris et al. (1958). This ensured listeners' perceptual attention was focussed on the changing vowel formant transitional information. To create a burst-less version of each token, the phone label of the initial consonant of the CVC (Consonant-Vowel-Consonant) sequence was removed before synthesis. All other labels were kept the same. This meant that the burst associated with the initial consonant closure did not appear in the synthesised token, but the formant transitions which relate to the consonant closure did appear, since these come under the vowel label. Pilot testing indicated that consonant contrasts created using the *Frm-DT* were not perceptually distinguishable by listeners, thus only the *Frm-GMM* system was used to create consonant stimuli for perceptual testing.

4.3. Participants and Procedure

All listeners were native speakers of English (28 female, 12 male) aged between 18 years and 38 years (average age: 23 years). All reported themselves as being free from speech or language disorders, hearing deficits, and histories of chronic otitis media.

All participants were tested individually in a sound-treated listening booth. Each participant listened to all three sets of speech continua: *Frm-GMM* consonants, *Frm-DT* vowels, and *Frm-GMM* vowels. Stimuli were presented over headphones at a comfortable listening level, in one ten minute session with two short breaks between stimulus sets. The listeners' task was to identify each CVC stimulus as one of the three members of the given contrast (e.g. "bet", "det" or "get" for the consonant stimuli) by clicking on the appropriate box in a graphical user interface.

Before each stimulus set, listeners were given the opportunity to listen to the two endpoints and the midpoint of the relevant continuum (i.e. the stimuli serving as the canonical examples of the possible responses for that set). Next, a practice test was administered to ensure that the listeners had understood the task. This practice test consisted of two repetitions each of the two endpoints and midpoint of the relevant continuum, presented in random order.

The stimuli within each set were presented five times each and in random order. Each listener heard a different randomisation. For all listeners, the consonant contrast stimuli were presented first. This was because, while the consonant continuum was burst-less, the vowel continua contained bursts. Presenting the consonant stimuli first meant that listeners avoided being influenced by exposure to the vowel stimuli to expect contrasts containing burst cues. Following presentation of the consonant stimuli, alternate listeners heard the *Frm-DT* stimuli followed by the *Frm-GMM* stimuli; the other half of the listeners heard the vowel sets in the opposite order.

4.4. Results

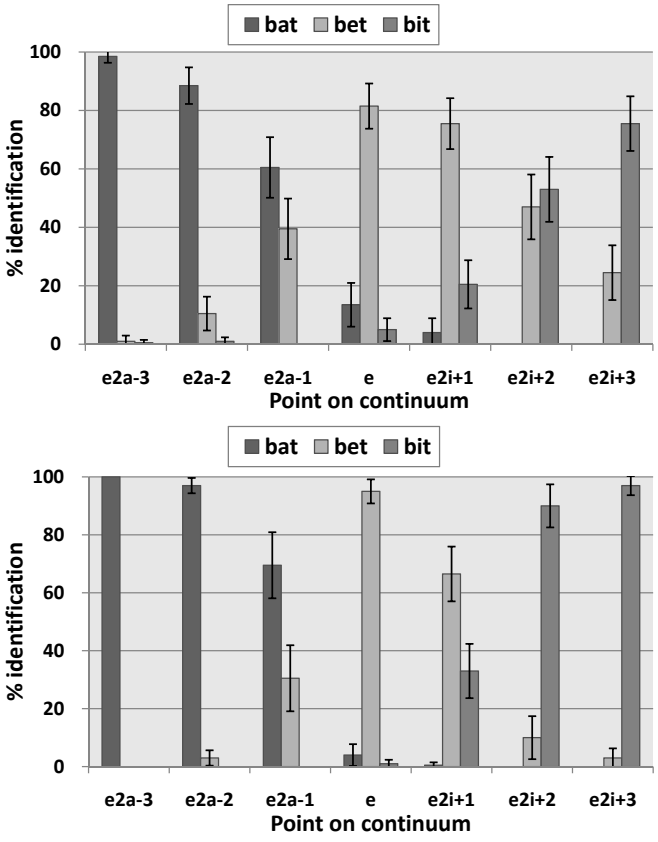

Figure 6: *Average listener responses to the vowel continuum produced by the* Frm-DT *system (top graph), and by the* Frm-GMM *system (bottom graph). In both graphs, the original, unmodified /*bɛt*/ token is marked as 'e'. The same token modified to contain vowel formant steady-state values appropriate for /*bæt*/ is marked as 'e2a-3', while the 'e' token modified to contain vowel formant steady-state values appropriate for /*bit*/ is marked as 'e2i+3'. Intermediate stimuli change in +/ − 50Hz steps between these three points (see text for details). Error bars represent 95% confidence interval.*

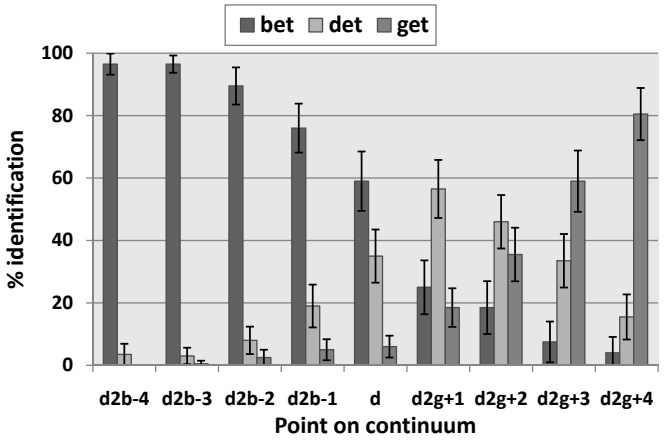

Figure 7: *Average listener responses to stimuli along the consonant contrast continuum. The original, unmodified /dɛt/ token is marked as 'd'. The same token resynthesised using the vowel formant onset contour from /bɛt/ is marked 'd2b-4'; 'd2g+4' indicates the same token resynthesised using the formant onset contour from /gɛt/. Intermediate stimuli change in equal steps between these three points (see text for details). Error bars represent 95% confidence interval.*

Fig. 6 shows the distribution of listener responses to the two sets of vowel stimuli. It is immediately apparent that three-formant stimuli more successfully model the differences between natural versions of these tokens than did the two-formant stimuli tested in Lei et al. (2011). The vowels produced by both the *Frm-GMM* system and the *Frm-DT* system achieved between 80% and 100% correct identification for the three end-/midpoint stimuli.

For the *Frm-GMM* continua, the three end- and midpoint stimuli received nearly 100% correct identification (token 'e2a-3', the /bæt/ endpoint, did in fact receive 100% correct responses). Additionally, the boundaries between the three phone categories are very steep, with little overlap. Contrast this with the responses to the *Frm-DT* stimuli, in which only the /bæt/ endpoint received over 80% correct responses, and in which the category boundaries are shallow with a great deal of overlap.

Fig. 7 shows the distributions of listener responses to the consonant contrast continuum. The first point to observe is that none of the end- or midpoint consonant stimuli ('d', 'd2b-4', and 'd2g+4'; see caption for details) received 100% "det", "bet" and "get" responses. Similarly, there is a certain amount of overlap at category boundaries. However, it is also clear that listeners were able to identify the stimuli as being from three different phone categories. The category that received the least listener agreement was "det". This is unsurprising given that these stimuli did not contain burst cues. Where identification is dependent solely on vowel onset formants, one should expect the least agreement for /dɛt/ rather than /bɛt/ or /gɛt/ as, of the three, /dɛt/ contains the least extensive transitional movement (Delattre et al., 1955). Note that HMM synthesised /dɛt/ tokens containing bursts consistently sound like "det".

5. Conclusion

This paper has described a formant-controllable HMM-based SPSS system. This system offers convenient, fine-grained control in terms of formants over both what is synthesised and exactly how it sounds. In this method, the conventional vocoder spectral features are modeled jointly with additional formant features within an HMM-based SPSS system. By introducing transform matrices to model the relationship between these two feature streams, the parameters of the distributions over the vocoder spectral features are made to depend upon the concurrent formant features. To perform synthesis, the formant features are predicted from the trained HMM first. These may then be manipulated at will and used in the prediction of the vocoder spectral features, from which the speech signal is resynthesized. In this way, changes to the formant features are reflected in the final synthesized speech.

Two experiments were designed according to prior phonetic knowledge in order to demonstrate the controllability of the proposed methods: vowel and consonant perception tests. Furthermore, two tying methods were evaluated here for grouping transform matrices: context-dependent tying and feature-space tying. On the basis of our results, we conclude that the context-dependent tying method is sufficient to offer control over vowel perception. However, the feature-space tying method gives superior control for the manipulation of both vowel and consonant perception. Overall, the proposed methods have been shown to synthesize speech effectively, while at the same time allowing prior knowledge of formants to be exploited and easily applied.

Acknowledgements

This work is partially funded by the National Nature Science Foundation of China (Grant No. 61273032) and the National Natural Science Foundation of China – Royal Society of Edinburgh Joint Project (Grant No. 61111130120). The research leading to these results was also partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 256230 (LISTA), and EPSRC grants EP/I027696/1 (Ultrax), EP/J002526/1 (CAF) and EP/I031022/1 (NST).

References

Acero, A., 1999. Formant analysis and synthesis using hidden Markov models, in: Proc. of EUROSPEECH, pp. 1047 – 1050.

Delattre, P.C., Liberman, A.M., Cooper, F.S., 1955. Acoustic loci and transitional cues for consonants. Journal of the Acoustical Society of America 27, 769–773.

Flanagan, J.L., Allen, J.B., Hasegawa-Johnson, M.A., 2008. Speech Analysis Synthesis and Perception. Springer-Verlag.

Harris, K.S., Hoffman, H.S., M., L.A., Delattre, P.C., Cooper, F.S., 1958. Effect of third-formant transitions on the perception of the voiced stop consonants. Journal of the Acoustical Society of America 30, 122–126.

Hawkins, S., Midgley, J., 2005. Formant frequencies of rp monophthongs in four age groups of speakers. Journal of the International Phonetic Association , 183–199.

Hu, H., Russell, M.J., 2010. Improved modelling of speech dynamics using non-linear formant trajectories for HMM-Based speech synthesis, in: Proc. of InterSpeech, pp. 821 – 824.

Jacewicz, E., Fox, R.A., Salmons, J., 2007. Vowel space areas across dialects and gender, in: Trouvain, J., Barry, W.J. (Eds.), the XVIth International Congress of Phonetic Sciences, pp. 1465–1468.

Kawahara, H., Morise, M., Banno, H., Takahashi, T., Nisimura, R., Irino, T., 2008. Spectral envelope recovery beyond the nyquist limit for high-quality manipulation of speech sounds, in: InterSpeech, pp. 650 – 653.

Kewley-Port, D., Zheng, Y., 1999. Vowel formant discrimination: Towards more ordinary listening conditions. The Journal of the Acoustical Society of America 106, 2945–2958.

Kienast, M., Sendlmeier, W.F., 2000. Acoustical analysis of spectral and temporal changes in emotional speech, in: ISCA Workshop on Speech and Emotion, pp. 92–97.

King, S., Karaiskos, V., 2010. The blizzard challenge 2010, in: Blizzard Challenge workshop.

Klatt, D., 1980. Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America 67, 971–995.

Ladefoged, P., Broadbent, D.E., 1957. Information conveyed by vowels. The Journal of the Acoustical Society of America 29, 98–104.

Lane, H., Tranel, B., 1971. The lombard sign and the role of hearing in speech. Journal of Speech and Hearing Research 14, 677–709.

Lei, M., Yamagishi, J., Richmond, K., Ling, Z., King, S., Dai, L., 2011. Formant-controlled HMM-based speech synthesis, in: Proceedings Interspeech 2011, pp. 2777–2780.

Ling, Z.H., Richmond, K., Yamagishi, J., 2011. Feature-space transform tying in unified acoustic-articulatory modelling for articulatory control of hmm-based speech synthesis, in: Proc. of InterSpeech, pp. 117–120.

Ling, Z.H., Richmond, K., Yamagishi, J., Wang, R.H., 2009. Integrating articulatory features into hmm-based parametric speech synthesis. Audio, Speech, and Language Processing, IEEE Transactions on 17, 1171 –1185.

Ling, Z.H., Wu, Y.J., Wang, Y.P., Qin, L., Wang, R.H., 2006. USTC system for blizzard challenge 2006 an improved hmm-based speech synthesis method, in: Blizzard Challenge workshop.

Liu, C., Kewley-Port, D., 2004. Straight: A new speech synthesizer for vowel formant discrimination. Acoustics Research Letters Online 5, 31–36.

Picart, B., Drugman, T., Dutoit, T., 2011. Analysis and synthesis of hypo and hyperarticulated speech, in: Speech Synthesis Workshop, pp. 270–275.

Shichiri, K., Sawabe, A., Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2002. Eigenvoices for hmm-based speech synthesis, in: Proc. of ICSLP, pp. 1269–1272.

Shinoda, K., Watanabe, T., 2000. Mdl-based context-dependent subword modeling for speech recognition. Acoustical Science and Technology 21, 79–86.

Sjolander, K., Beskow, J., Gustafson, J., Lewin, E., Carlson, R., Granstrom, B., 1998. Web-based educational tools for speech technology, in: International Conference on Spoken Language Processing, pp. 3217–3220.

Tachibana, M., Yamagishi, J., Masuko, T., Takao, K., 2005. Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. IEICE TRANSACTIONS on Information and Systems E88-D, 2484–2491.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speech parameter generation algorithms for hmm-based speech synthesis, in: Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on, pp. 1315 –1318 vol.3.

Yamagishi, J., Kobayashi, T., 2007. Average-voice-based speech synthesis using hsmm-based speaker adaptation and adaptive training. IEICE TRANSACTIONS on Information and Systems E90-D, 533–543.

Zen, H., Tokuda, K., Black, A.W., 2009. Review: Statistical parametric speech synthesis. Speech Communication 51, 1039–1064.