Formant-based Frequency Warping for Improving Speaker Adaptation in HMM TTS

Xin Zhuang\textsuperscript{1,2}, Yao Qian\textsuperscript{1}, Frank Soong\textsuperscript{1}, Yijian Wu\textsuperscript{3}, Bo Zhang\textsuperscript{2}

\textsuperscript{1} Microsoft Research Asia, Beijing, China
\textsuperscript{2} College of Software, NanKai University, Tianjin, China
\textsuperscript{3} Microsoft China, Beijing, China

\texttt{(yaoqian,frankkps,yijiwu@microsoft.com)}

Abstract

Vocal Tract Length Normalization (VTLN), usually implemented as a frequency warping procedure (e.g., bilinear transformation), has been used successfully to adapt the spectral characteristics to a target speaker in speech recognition. In this study, we exploit the same concept of frequency warping but concentrate explicitly on mapping the first four formant frequencies of long vowels from source and target speakers. A universal warping function is thus constructed for improving MLLR-based speaker adaptation performance in TTS. The function first warps the frequency scale of the source speaker’s speech data toward that of the target speaker and an HMM of the warped features is trained. Finally, MLLR-based speaker adaptation is applied to the trained HMM for synthesizing the target speaker’s speech. When tested on a database of 4,000 sentences (source speaker) and 100 sentences of a male and a female speaker (target speakers), the formant-based frequency warping has been found very effective in reducing the objective, log spectral distortion over the system without formant frequency warping. The improvement is also subjectively confirmed in AB preference and ABX speaker similarity listening tests.

Index Terms: speech synthesis, speaker adaptation, frequency warping

1. Introduction

The state-of-the-art, speaker trained Text-to-Speech (TTS) system, can deliver good quality, synthesized speech for general applications. However, applications which require more customized voices pose a challenge of how to convert a well-trained single speaker TTS to a new speaker with only limited training data. Voice conversion/adaptation can efficiently change the voice of a speaker or its characteristics into that of a different speaker in a natural and seamless way.

In recent years, the voice quality and similarity of speech generated by voice conversion and speaker adaptation have been largely improved by various statistical approaches. One of the most popular approaches to voice conversion is a probabilistic conversion based upon a Gaussian mixture model (GMM) \cite{1,2}, which uses parallel utterance pairs of the source and target speakers to train a GMM for converting the speech from a source speaker to a target in a minimum-mean-square error (MMSE) or maximum likelihood (ML) sense. Speaker adaptation, which transforms the model of a source speaker to match that of a target speaker, is developed for automatic speech recognition (ASR) application. Since the HMM-based speech synthesis was proposed, the speaker adaptation techniques, e.g., maximum a posterior (MAP), maximum likelihood linear regression (MLLR) and speaker adaptive training (SAT), also have been applied to change the voice characteristics of synthesized speech \cite{3,4}. A new adaptation algorithm called constrained structural maximum a posteriori linear regression (CSMAPLR) is also proposed in \cite{4}. Eigenvoice is another GMM-based voice conversion method \cite{5}.

ASR systems generally use vocal tract length normalization (VTLN) as a preprocessing step prior to speaker adaptation, e.g., MAP and MLLR, to remove individual speaker characteristics. VTLN equalizes the speaker difference by normalizing the vocal tract length. There are two problems involved in VTLN: 1) how to obtain the vocal tract length (VTL); instead of measuring it directly, usually we use a warping or normalization factor to represent the difference in VTL; 2) given a warping or normalization factor, how to normalize the length. It is generally implemented with a warping frequency axis in extracting acoustic feature vectors.

Many attempts in voice conversion and speaker adaptation have been tried. In \cite{6}, VTLN based speaker adaptation is tried in statistical speech synthesis which yields additional improvement over CMLLR adaptation. The warping functions can be linear, piecewise linear, bi-linear or nonlinear. A bilinear warping estimated from very little data is used in \cite{6}. The warping factors are conventionally obtained by grid search based on the ML criterion. It finds the optimal warping factor in the ML criterion, which is consistent with HMM training. However, it is relatively expensive computationally. ML criterion does not have a direct relation with voice transformation performance in speaker similarity. VTL has relationship with formant positions, which are inversely proportional to the VTL, and hence the warping factor can be obtained from formant estimation. The frequency warping function generated by mapping formants of the source and target speakers is employed to voice conversion in \cite{7}. Formant based approach is more straightforward and less expensive in computation. However, we still need to be careful in getting robust formant estimates of both the source and target speakers.

In this paper, we adopt a formant-based frequency warping approach to match the perceptually sensitive spectral peaks between source and target speakers. The frequency warping function is generated by mapping the first four formant pairs extracted from the same long vowels of source and target speakers. It is a more direct and perceptually meaningful to normalize the center frequencies of formants caused by the speaker-dependent vocal tract length than other VTLN methods. Experimental results also demonstrate its effectiveness, both objectively and subjectively.

The rest of paper is organized as follows. In Section 2, we give an overview of HMM-based speech synthesis and adaptation. The formant-based frequency warp and corresponding experimental results and analysis are presented in Sections 3 and 4, separately. We draw our conclusions in Section 5.
2. The Overview of HMM-based Speech Synthesis and Adaptation

In conventional HMM-based speech synthesis system, the spectral envelope, fundamental frequency, and duration are modeled simultaneously by the corresponding HMMs. In synthesis, for a given text sequence, speech parameter trajectories and corresponding signals are then generated from the trained HMMs in the Maximum Likelihood (ML) sense. Recently, hidden semi-Markov model (HSMM), in which the duration of each state is modeled explicitly and trained by the actual collected duration, is implemented in HTS version 2.1 [8] for training, synthesis and adaptation. In HSMM based system, spectrum and prosody of speech are modeled simultaneously by context dependent HSMMs, and speech parameter vector sequences are generated from the HSMMs themselves [9]. The explicit state duration distributions can be adapted together with the state output distributions [10].

The schematic diagram of our HSMM-based speech synthesis and adaption system is shown in Fig. 1. We use line spectrum pair (LSP) [11] as the spectrum feature in HSMM training. LSP parameter has a good interpolation property and correlates well with the “formants” or spectral peaks. Spectral envelopes are first estimated by STRAIGHT [12] and LPC then converted LSPs. For speaker adaptation, we first do frequency warping of the spectrum and adjustment of F0 toward the target speaker. The details of spectrum warping and F0 adjustment will be introduced in Section 3. We use LSPs converted from warped spectrum and modified F0 to train HSMMs. The HSMMs thus trained are closer to the target speaker than the models trained by using the original source speaker’s features. MLLR is then used to adapt the model trained by warped features. In order to get better performance with limited adaptation data, we conduct a global transform to the model firstly, i.e., all states of the models share the same transform matrix. Then more transform matrices are employed further.

Fig. 1. The schematic diagram of our HSMM-based speech synthesis and adaption system

3. Formant-based Frequency Warping

We adopted the formant-based frequency warping proposed by Shuang [7] to estimate the warping function. The warping function is piecewise linear derived by mapping formant frequencies of matched vowel pairs from source to target speakers.

We choose the formant pairs of long vowels to represent the differences between speakers. We align vowel segments of the same word uttered by source and target speakers. Then, we select the stationary segments in the aligned vowels, where expect the formants exhibit fairly stationary behavior. The length of the segment is set to 40ms and the formant parameter is averaged over the whole segment.

The formant parameters are extracted automatically by a speech signal process tool “wavesurfer” [13], which is an open source tool for sound visualization and manipulation, automatic formant extraction, etc. Formant errors are manually corrected. When more than one vowel is used, multiple pairs of the mapping are established.

The first four formants of the selected stationary vowel segments are used to represent one speaker’s formant characteristics. To define a piecewise linear frequency warping function from the source speaker to the target speaker, we need to get some key mapping pairs firstly. The four pairs of mapped formants, \([F_1^s, F_1^t], \ldots, [F_4^s, F_4^t]\), are used as four key anchoring points in addition to the points \([0, 6000], [8000, 8000]\) as the first and the last ones. Linear interpolation is used to generate mappings between two adjacent anchoring points. The warping function generation with the piece-wise linear function derived from mapped formants is illustrated in Fig. 2.

The warping function is based upon the points derived from mapped formant frequencies.

We use the frequency warping function to warp the frequencies of spectrum according to Eq. 1.

\[
\hat{S}(w) = s(f(w))
\]  

Fig. 2. The piecewise linear warping function is based upon the points derived from mapped formant frequencies.

\[
F_0 = \frac{(F_0^t - u_s)}{\sigma_s}, \sigma_t + u_t
\]
4. Experiments and Results

4.1. Experimental Setup

A phonetically and prosodically rich corpus in broadcast news style is used as the source speaker data in our experiments. It is a U.S. English corpus recorded by a male speaker. The training data consists of 4,000 sentences around 5 hours. Two other speakers, each recorded 100 sentences, are used as target speakers in our adaptation experiments. The first speaker is a female one and her 100 sentences are phonetically balanced. The second speaker is a male one and his data is downloaded from internet. His recording is slightly noisy and SNR is around 20 dB SNR. Ten sentences of each speaker are used as testing data. We use the symbols in labeling the gender and source/target speaker, e.g. M_S, F_T and M_T denote male source, female target and male target speakers.

Speech signals are sampled at 16 kHz, windowed by a 25-ms window with a 5-ms shift. Spectral envelopes are estimated by STRAIGHT [12] and warped by our formant-based approach, as mentioned in Section 3. The 40th order LPC coefficients are transformed into static LSPs and their dynamic counterparts. Five-state, left-to-right HSMM phone models, where each state is modeled with a single Gaussian, diagonal covariance output distribution, are adopted. The phonetic and prosodic contexts are used as question set in growing the decision tree for state clustering. We conduct a global adaptation firstly, and then use more transform matrices to perform further speaker adaptation.

4.2. Experimental Results and Analysis

We choose 5 long vowels embedded in 10 sentences of the source and target speakers to extract formants for the warping function estimation. The vowels and corresponding words are listed in Table 1. The vowel pair(s) of source and target speakers can be used to generate a frequency warping function for each vowel. We then average the five such functions to obtain a universal warping function for warping the frequency of a source speaker toward that of the target speaker. The corresponding two frequency warping functions generated for transforming from M_S to F_T and M_S to M_T are shown in Fig. 3.

Table 1. Vowels and corresponding words selected for formant extraction.

<table>
<thead>
<tr>
<th>Word</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>choose</td>
<td>/ʃ</td>
</tr>
<tr>
<td>need</td>
<td>/n\</td>
</tr>
<tr>
<td>force</td>
<td>/f\</td>
</tr>
<tr>
<td>Hard</td>
<td>/\</td>
</tr>
<tr>
<td>heard</td>
<td>/\</td>
</tr>
</tbody>
</table>

Objective and subjective measures are used to evaluate the performance of our approach in testing data. Adaptation quality is measured objectively in terms of distortions between target original speech and speech synthesized by adapted source speaker’s model. Since the predicted HSMM state durations of the generated utterances are in general not the same as those of the recorded speech, state durations of the original recording are obtained by forced alignment and used as given duration in synthesis. Both spectrum and pitch can then be compared on a frame-synchronous basis between the original and synthesized target utterances.

Table 2 shows the log spectrum distance (LSD) and room mean square error (RMSE) of F0 for the test sentences of original target speaker and generated by HSMM trained by the warped feature of source speaker. It can be seen that the models trained by using warped features are closer to the target speakers. The LSD improvements are more effective in adapting from M_S to F_T than M_S to M_T. The LSDs are 1.57 dB (9.01 vs 7.44) and 0.24 dB (8.49 vs 8.25), respectively. Considering the difference of vocal tract length between female and male speakers is larger than that between male and male speakers, the warping function for cross-gender adaptation should be more effective than intra-gender adaptation. The F0 adjustment shows a similar trend.

<table>
<thead>
<tr>
<th>Word</th>
<th>M_S Model</th>
<th>M_S Model with Warped Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>speaker</td>
<td>F_T</td>
<td>M_T</td>
</tr>
<tr>
<td>Spectrum (dB)</td>
<td>9.01</td>
<td>8.49</td>
</tr>
<tr>
<td>F0 (Hz)</td>
<td>95.47</td>
<td>23.87</td>
</tr>
</tbody>
</table>

Fig. 3. Two frequency warping functions generated to transform M_S to F_T and M_S to M_T.

The HSMMs, which are trained with the warped features of source speaker, are further adapted by MLLR with 5, 10, 50 and 100 sentences of the target speakers. Fig. 4 shows the corresponding LSD of the testing sentences of the original target speaker and generated by the adapted HSMMs. The proposed formant-based warping function is effective in reducing the LSD. The two performance curves always show the frequency warping is always better than without. The improvement of LSD by frequency warping in M_S to F_T conversion, i.e., there are 0.3 dB and 0.2 dB with 5 sentences and 100 sentences, respectively. Similarly, in M_S to M_T conversion, i.e., 1.0 dB and 0.2 dB by using 5 sentences and 100 sentences, respectively. However, by comparing with the two sets of curves (top and bottom) in Fig. 4, we observe that the absolute LSD is lower for cross-gender adaptation than same-gender one. This may due to the fact that the adaptation sentences of M_T, collected from internet, are slightly noisy, reverberant and in a different speaking (i.e., conversational) style. Formant extraction from it may not be as accurate and result in less effective frequency warping function from M_S to M_T. The RMSE of F0 and duration of the test sentences between original target speaker and generated by HSMMs are also decreased with the increasing number of adaptation sentences, but they are not so obviously as the improvement of LSD.

The effectiveness of frequency warping is further measured subjectively by two listening tests. One is an AB preference test, in which subjects select the preferred one from a pair of sentences in term of naturalness. The other test is an ABX similarity test, which measures the perceptual distance from source to target speaker. X is the original sentences from target speakers. Subjects were asked to select between A and B...
closer (in speaker similarity) to sentence X. Six subjects participated in the listening tests. Fig. 5 and 6 show the results of AB and ABX tests for the synthesized sentences by HSMMs adapted by 50 sentences with/without feature warping. The results of listening tests show that our formant-based frequency warping approach can significantly improve the adaptation performance in both synthesis naturalness and speaker similarity.

5. Conclusions

In this paper, we use a formant-based frequency warping between source and target speakers as a VTLN preprocessing procedure for HMM-based TTS applications. After the frequency warping preprocessing, HSMM phone models are trained with LSPs estimated from the warped spectrum and F0 are adjusted according to the distribution of the target speaker. Finally MLLR is employed to adapt HSMMs. Experimental results show that our frequency warping approach result in an additional improvement over the MLLR adaptation, both objectively and subjectively.

6. References