VOICE CHARACTERISTICS CONVERSION FOR TTS USING REVERSE VTLN

Matthias Eichner, Matthias Wolff and Rüdiger Hoffmann

Dresden University of Technology, Germany

ABSTRACT

In the past, several approaches have been proposed for voice conversion in TTS systems. Mostly, conversion is done by modification of the spectral properties and pitch to match a certain target voice. This conversion causes distortions that deteriorate the quality of the synthesized speech. In this paper we investigate a very simple and straightforward method for voice conversion. It generates a new voice from the source speaker instead of generating a certain target speaker's voice. For application in TTS systems it is often sufficient to synthesize new voices that sound sufficiently different to be distinguishable from each other. This is done by applying a spectral warping technique that is commonly used for speaker normalization in speech recognition systems called vocal tract length normalization (VTLN). Due to the low requirements of resources this method is especially suited for embedded systems.

1. INTRODUCTION

Voice conversion is used in TTS to adapt the synthesized speech to a certain target speaker. This offers the possibility for personalized speech synthesis, where the synthesizer is able to speak with any desired voice. Conversion is mostly done by modification the spectral properties and pitch [5]. The signal processing necessary to realize the spectral adaptation causes distortions that lead to a deterioration of the overall quality. In this paper we investigate a method for voice conversion that generates a new voice from the source speaker while introducing as few distortions as possible. In contrast to the mentioned approaches it does not try to produce the voice of a given target speaker. The conversion is done by applying a spectral warping technique that is commonly used for speaker normalization in speech recognition systems called vocal tract length normalization (VTLN) [15].

The motivation for this work has two aspects. Firstly, it is motivated by our work on an unified approach for speech synthesis and recognition (UASR) [1]. The goal of this approach is to design a system that uses the same databases and algorithms for both, speech synthesis and recognition. It is straightforward thinking to reverse normalization techniques used in the recognition branch by de-normalization in the synthesis part. The combination of normalization and de-normalization solves an ostensible conflict in integrating speech synthesis and recognition: recognizers shall be speaker independent, where speech synthesizers shall speak with a voice having certain characteristics. In our system, the acoustic models describe the average voice derived from speaker normalized training data. In the synthesis step an arbitrary voice can be synthesized from average voice by applying de-normalization techniques.

The second motivation for this work is the generation of new voices for our concatenative TTS system *Dress* [4]. The design and production of new signal databases is an expensive and time consuming process. Therefore we were looking for a way to build new voices from existing databases. For TTS systems with a small footprint in embedded environments this method is especially interesting, because the conversion could be moved entirely in the synthesis stage. By applying this technique, different voices can be synthesized using a single speaker database. As we will show in this paper, the quality of the signal is not affected by the additional processing.

In the following we will describe the vocal tract length normalization method we are using in the recognition part, explain how this normalization is reversed for speech synthesis and will discuss first experimental result.

2. VTLN IN SPEECH RECOGNITION

The use of VTLN in speech recognition has been widely investigated ([3][6][7][10][15]). In this section, we will briefly describe the concept and how we use VTLN in our own system. The variance of the vocal tract length of speakers is, among others, one major reason for the inter-speaker variability. This variation leads to different locations of the formant peaks for the same utterance spoken by different speakers. Longer vocal tracts result in a compression of the frequency axis, where shorter vocal

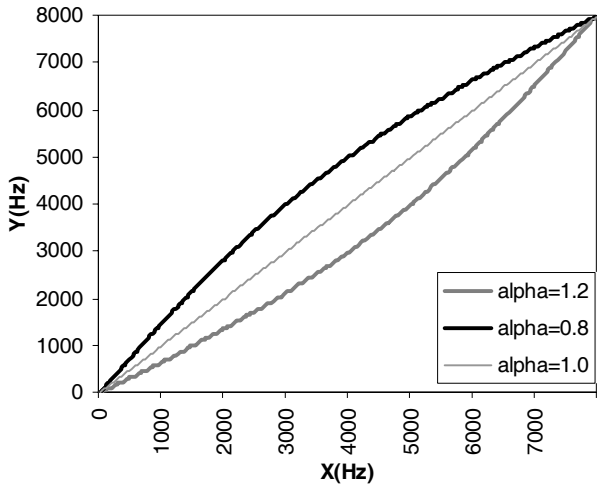

Figure 1: Bilinear warping function for different values of the warping factor α

tracts stretch the formant patterns. To build speech recognizers as robust as possible, a normalization of feature data is applied to eliminate the VTL variation. This normalization is done by warping the signal in the frequency domain to match the acoustic models of the recognizer, hence to move the formant positions in direction of the average speaker. There have been several methods proposed how to estimate the warping factor and what kind of warping functions can be applied. In our experiments we have used the bilinear warping function $\varphi_\alpha(\omega)$:

$$\varphi_\alpha(\omega) = \omega + 2\arctan\left(\frac{(1-\alpha)\sin(\omega)}{1-(1-\alpha)\cos(\omega)}\right) \quad (1)$$

Where α is the speaker-specific warping factor. A factor α>1.0 results in compression of the spectrum, whereas α<1.0 corresponds to stretching the spectrum (figure 1). The warping factors for the speakers in our database were estimated by line search. For every speaker a factor α was selected in a range of $\alpha - \delta$ and $\alpha + \delta$, where the phoneme recognition rate was maximal. The obtained set of factors was used to train a new set of acoustic models. The search for the best α and the following retraining was repeated until the change of factors between two iterations fell below a certain threshold.

3. REVERSE VTLN IN SPEECH SYNTHESIS

VTL de-normalization warps the signal from average voice back to a voice having definite male or female characteristics. As mentioned above, this is useful for parametric speech synthesizers ([2][12][13]) which model the speech signal using statistic models trained using data from different speakers. These models represent the average voice and there is some processing necessary to synthesize a voice having certain characteristics from average models. Another approach that has been proposed is to adapt the acoustic models to the target speaker [11]. In contrast, the reverse VTLN leaves the acoustic models unchanged, but modifies the extracted parameters during synthesis. The combination of both approaches, speaker normalization and speaker adaptation, is advantageous in speech recognition [16] and could improve the quality of parametric TTS systems as well.

In our system we use cepstral features and a synthesis filter that reconstructs the speech signal from cepstral coefficients. The input signal is segmented (length 24ms, continuation rate 10ms) and windowed using blackman window. The real cepstrum is computed as:

$$c_n = F^{-1}\left\{\left|\ln\left(F\left\{s(n)\right\}\right)\right|\right\} \quad (2)$$

where F denotes the N points discrete Fourier series, $s(n)$ the windows speech frame and c_n are the N cepstral coefficients. To recover the speech signal out of the coefficients we use a digital filter. Since the input signal for the filter is an impulse sequence for voiced segments and noise for unvoiced segments, the transfer function of such a filter is given by:

$$H(\omega) = \left|F\left\{s(n)\right\}\right| \quad (3)$$

From (2) follows

$$F\{c_n\} = \left|\ln\left(F\left\{s(n)\right\}\right)\right| \quad (4)$$

$$e^{F\{c_n\}} = H(\omega) \quad (5)$$

where $F\{c_n\}$ is the transfer function of the digital filter used to approximate the absolute value of the natural logarithm of the speech frame's spectrum. The exponential function is approximated using Páde approximation. The resulting synthesis filter is a composite filter consisting of a FIR filter that realizes the transfer function and an IIR filter which approximates the exponential function [14].

4. DATABASE

For our experiments we used the German PhonDAT II database [9]. It consists 200 sentences read by 16 speakers. The 3200 signal files were recorded in high quality 16-bit, 16 kHz. 6 out of 16 speakers were female; the remaining 10 speakers were male. We extracted the pitch information for every file in database using a wavelet-based pitch tracker. The pitch information includes the voiced/unvoiced information. For every speaker in the database we have determined the speaker-specific warping factor α using the described line search.

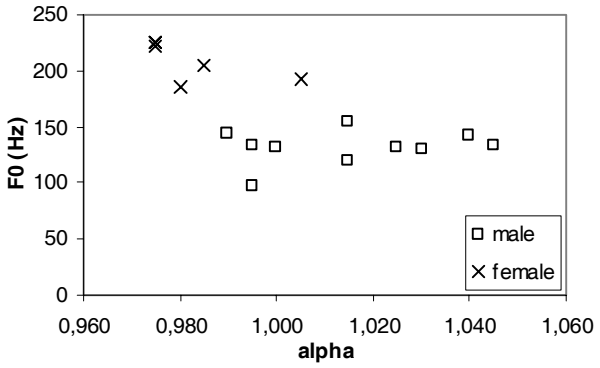

Figure 2: *F0/α* plot for the speakers from the German PhonDAT II database

5. EXPERIMENTS

To evaluate suitability of reverse VTLN for speech synthesis we have conducted two re-synthesis experiments. Re-synthesis allows an objective judgment of the quality loss in reconstructed signal by avoiding the additional processing stages in the TTS system. The first goal was to answer the question, if additional signal processing introduced by warping the spectra affects the quality of the synthesized signal. Therefore we performed a mean opinion score (MOS) test using samples of recorded speech, re-synthesized speech and warped speech. The second objective was to find out, if the warped signals have significantly different characteristics to be distinguishable from the original speaker.

The utterances were chosen from the database by randomly selecting 5 sentences from each of eight different speakers (4 male and 4 female). One sample per speaker remained unchanged, one sample was re-synthesized using cepstral synthesis and 3 samples were warped to produce different target voices according to table 1.

Speaker	**Original**	**W 1**	**W 2**	**W 3**
AWE	Female *F0*=195 α=0.98	Female *F0*=240 α=1.05	Male *F0*=140 α=0.95	Male *F0*=120 α=0.9
CHK	Male *F0*=138 α=1.015	Male *F0*=120 α=0.95	Female *F0*=180 α=1.05	Female *F0*=220 α=1.1
CSC	Male *F0*=144 α=0.99	Male *F0*=120 α=0.95	Female *F0*=180 α=1.05	Female *F0*=220 α=1.1
KMA	Female *F0*=235 α=0.975	Female *F0*=240 α=1.05	Male *F0*=140 α=0.95	Male *F0*=120 α=0.9
MKN	Female *F0*=213 α=0.975	Female *F0*=240 α=1.05	Male *F0*=140 α=0.95	Male *F0*=120 α=0.9
RTD	Female *F0*=229 α=0.975	Female *F0*=240 α=1.05	Male *F0*=140 α=0.95	Male *F0*=120 α=0.9
SAT	Male *F0*=154 α=1.015	Male *F0*=120 α=0.95	Female *F0*=180 α=1.05	Female *F0*=220 α=1.1
TPO	Male *F0*=133 α=1.025	Male *F0*=120 α=0.95	Female *F0*=180 α=1.05	Female *F0*=220 α=1.1

Table 1: Experimental settings for warped samples

The warping factors listed in column 2 in table 1 are the factors estimated using line search (see section 2). They describe the amount of warping necessary to approximate the source voice to the average voice. In the synthesis experiments the warping was performed in the opposite direction. In this case a warping factor less one means generation of a male voice characteristic, a warping factor greater one stands for generation of female voice characteristics, respectively. The excitation signal for re-synthesis was generated by re-sampling of the original pitch to match the target fundamental frequency.

12 subjects participated in the listening tests. In the MOS test they were asked to rate the presented speech samples on a scale between 1 and 5. Figure 3 shows that the speech quality of the re-synthesized signal is rated about 1.5 points below the original signal.

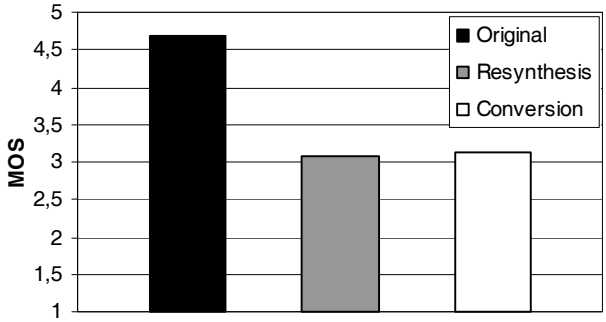

Figure 3: MOS rating of natural speech, re-synthesized speech samples and warped speech

A possible explanation for this low rating is that the test did not include samples generated by a TTS system. The interesting and important fact of this test is that the frequency warping did not cause additionally distortion. In the second test 16 pairs of speech samples were presented. The listeners were asked to decide whether the two sentences were uttered by the same speaker or not.

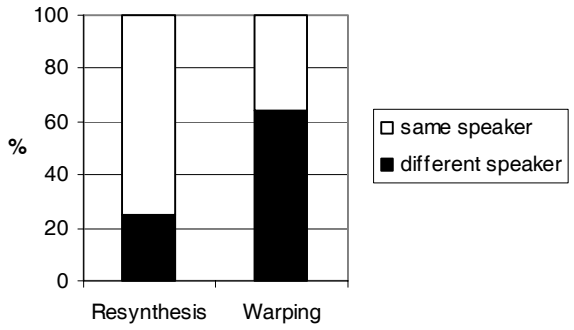

Figure 4: Listeners were asked to decide whether two sentences were spoken by the same speaker or not.

The samples included natural speech, re-synthesized speech and warped speech samples. The new voices generated using reverse VTLN were hardly recognized from the source voices by the listeners (figure 4). Surprisingly, even the comparison between natural speech and re-synthesized signal did not yield 100%.

6. CONCLUSION

The reverse VTLN can be used in speech synthesis to generate different voices from a single voice, either represented by a unit database of concatenative synthesizers or by acoustic models of a parametric TTS system. The listening tests showed, that the quality is not affected by the additional processing and that the method is able to produce distinguishable voices from a single speaker. The low computational requirement qualifies the method especially for application in embedded systems. In our unified system for speech recognition and synthesis the approach enables us to synthesize speech from average voice models having certain characteristics.

7. REFERENCES

[1] Eichner, M., Wolff, M., Hoffmann, R., “A unified approach for speech synthesis and speech recognition using Stochastic Markov Graphs”, Proc. ICSLP, Beijing (China), vol. 1, pp. 701-704, 2000.

[2] M. Eichner et al., “Speech synthesis using stochastic Markov graphs,” Proc. ICASSP, Salt Lake City (USA), 2:829-832, 2001.

[3] E. Eide and H. Gish, "A Parametric Approach to Vocal Tract Length Normalization," In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 346-348, 1996.

[4] R. Hoffmann et al., "A multilingual TTS system with less than 1 MByte footprint for embedded applications", Proc. ICASSP, Hong Kong, 2003.

[5] Kain and M. Macon, “Spectral voice conversion for text-to-speech synthesis,” Proc. ICASSP, 1:285-288, 1998.

[6] L. Lee and R. Rose, "A Frequency Warping Approach to Speaker Normalization," IEEE Transactions on Speech and Audio Processing, vol. 6, pp. 49-60, Jan. 1998.

[7] J. McDonough, W. Byrne, and X. Luo, "Speaker Normalization with All-Pass Transforms," In Proceedings of the International Conference on Spoken Language Processing, vol. 6, pp. 2307-2310, 1998.

[8] M. Ostendorf, I. Bulyko, “The impact of speech recognition on speech synthesis”,TTSWS,2002.

[9] H.R. Pfitzinger, "10 Years of PhonDat-II: A Reassessment." Proc. ICSLP, Denver, vol. 1, pp. 369-372, 2002.

[10] D. Pye and P. C. Woodland, "Experiments in Speaker Normalization and Adaptation for Large Vocabulary Speech Recognition", In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 1047-1050, 1997.

[11] M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi, Speaker adaptation for HMM-based speech synthesis system using MLLR,” Proc. ESCA/COCOSDA Workshop on Speech Synthesis, 273-276, 1998.

[12] M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi, “Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR,” Proc. ICASSP, 2:805-808, 2001.

[13] Z. Tychtl, and J. Psutka, “Speech production based on the mel-frequency cepstral coefficients”, Proc. EUROSPEECH, Budapest, vol. 5, pp. 2335-2338, 1999.

[14] R. Vích, Z. Smekal, “New method of composite FIR and IIR filtering for cepstral speech synthesis.”, Proc. 17th IASTED Internat. Conf. Applied Informatics, Innsbruck, Austria, pp. 264-267. 1999.

[15] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker Normalization on Conversational Telephone Speech," In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 339-341, 1996.

[16] P. Zhan, M. Westphal, M. Finke, and A. Waibel, “Speaker Normalization and Speaker Adaptation - a Combination for Conversational Speech Recognition”. Proceedings of Eurospeech Conference, Greece, 1997.

[17] P. Zhan and A. Waibel, "Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition", CMU-CS-97-148, Carnegie Mellon University, Pittsburgh, PA, May 1997.