A STUDY ON RESIDUAL PREDICTION TECHNIQUES FOR VOICE CONVERSION

David Sündermann, Antonio Bonafonte

Universitat Politècnica de Catalunya
Department of Signal Theory and Communications
C/ Jordi Girona, 1 i 3, 08034 Barcelona, Spain
{suendermann,antonio}@gps.tsc.upc.es

Hermann Ney

RWTH Aachen – University of Technology
Computer Science Department
Ahornstr. 55, 52056 Aachen, Germany
ney@cs.rwth-aachen.de

Harald Höge

Siemens AG
Corporate Technology
Otto-Hahn-Ring 6, 81739 Munich, Germany
harald.hoege@siemens.com

ABSTRACT

Several well-studied voice conversion techniques use line spectral frequencies as features to represent the spectral envelopes of the processed speech frames. In order to return to the time domain, these features are converted to linear predictive coefficients that serve as coefficients of a filter applied to an unknown residual signal. In this study, we compare several residual prediction approaches that have already been proposed in the literature dealing with voice conversion. We also present a novel technique that outperforms the others in terms of voice conversion performance and sound quality.

1. INTRODUCTION

Voice conversion is the adaptation of the characteristics of a source speaker's voice to those of a target speaker [1]. Over the last few years, the interest in voice conversion has risen significantly. This is due to its application to the individualization of text-to-speech systems, whose voices, in general, have to be created in a rather time-consuming way requiring human assistance [2].

The most popular voice conversion technique is the application of a linear transformation to the spectra of speech frames [3]. The transformation parameters are estimated using a Gaussian mixture model to describe the characteristics of the considered speech data. In doing so, we cannot use the full spectra of the processed time frames, as their high dimensionality leads to parameter estimation problems. Therefore, the spectra are converted to features that are linearly transformed and then converted back to the frequency domain or directly to the time domain.

After investigating several feature representations, we were able to confirm the statement of H. Ye and S. Young [4] that using line spectral frequencies results in much better sound quality than other feature types as for instance mel frequency cepstral coefficients. However, line spectral frequencies represent the spectral envelope in a rather smooth way tracing the shape of the formants and antiformants. Thus, spectral details that are typical for natural speech are lost, and the result sounds rather synthetic. To overcome the weakness of the signal representation by line spectral frequencies, L. M. Arslan and D. Talkin [5] proposed to predict the target residuals that belong to the converted speech. These residuals are then filtered based on linear predictive coefficients that are derived from the line spectral frequencies. Finally, the spectral details of the residuals result in more natural spectra of the converted speech, a consideration that led to the term *high resolution* voice conversion [6].

The problem now is how to predict the target residuals of the converted speech. In the following section, we briefly describe three solutions to that problem that have already been proposed in the literature and, in Section 3, we derive a new technique that is simpler and more general than the others. Then, we examine these approaches using a Spanish cross-gender corpus by means of listening tests in Section 4. The results of this evaluation are discussed in Section 5. It turns out that the novel approach outperforms the others in terms of voice conversion performance and sound quality.

2. RESIDUAL PREDICTION: RELATED WORK

2.1. The Trivial Solution: Copying Residuals

Let us suppose that the vocal tract characteristics of a speaker's speech are represented by line spectral frequencies, and the residuals correspond to the excitation signal that hardly contains speaker-dependent information. Then, the simplest idea for the residual prediction is to take the residuals of the source speech and filter them by means of the converted features. This technique was used by A. Kain and M. W. Ma-

con, but they stated that "merely changing the spectrum is not sufficient for changing the speaker identity" [7]. Most of the listeners "had the impression that a 'third' speaker was created".

Hence, it seems that the residuals contain a lot of speaker-dependent information so that the contribution of the spectral transformation to the properties of the converted signal can hardly be distinguished from that of the residual prediction. To separate both contributions, Duxans et al. [8] extracted the residuals from the reference (target) speech that was aligned to the source speech and restricted their investigations to the spectral conversion. Utilizing the reference residuals as excitation signal should result in the best voice conversion performance in terms of sound quality and the ability for identity conversion compared to an arbitrary residual prediction technique. Therefore, in this paper, we want to use this method as a standard of comparison.

2.2. Residual Codebook Method

In addition to the observation that the residual signal also contains speaker-dependent information, it has to be mentioned that the line spectral frequencies which describe the vocal tract characteristics and the corresponding residuals are even correlated. This insight led to the idea that the residuals of the converted speech could be predicted based on the converted feature vectors and resulted in the following residual prediction technique [6].

In training phase, for each pitch-synchronous frame, we compute the linear predictive coefficients and convert them to a cepstral representation. Then, the probability distribution of the set of all linear predictive cepstral vectors seen in training is modeled by means of a Gaussian mixture model. Now, we determine the typical residual magnitude spectra $\hat{m}_i$ for each mixture component i by computing a weighted average of all residual magnitude spectra m_n seen in training where the weights are the posterior probabilities $p(i|v_n)$ that a given cepstral vector v_n belongs to the mixture component i:

$$\hat{m}_i = \frac{\sum_{n=1}^{N} m_n p(i|v_n)}{\sum_{\nu=1}^{N} p(i|v_\nu)} \; .$$

The mixture-dependent phase spectra are taken from the centroids of each mixture component in the following way:

$$\hat{\varphi}_i = \varphi_{\hat{n}} \quad \text{with} \quad \hat{n} = \underset{n\,=\,1,\ldots,N}{\arg\max}\, p(i|v_n) \; .$$

During operation phase, for each frame, we obtain a converted cepstral vector $\tilde{v}$ that serves as basis for the prediction of the residual magnitude spectrum $\tilde{m}$ by calculating a weighted sum over all mixture components:

$$\tilde{m} = \sum_{i=1}^{I} \hat{m}_i p(i|\tilde{v}) \; .$$

By selecting the most likely mixture component, we can derive the required phase spectrum:

$$\tilde{\varphi} = \hat{\varphi}_{\tilde{\imath}} \quad \text{with} \quad \tilde{\imath} = \underset{i\,=\,1,\ldots,I}{\arg\max}\, p(i|\tilde{v}) \; .$$

To avoid artifacts due to the discreteness of the phase spectra, the trajectories of each harmonic phase are smoothed by zero-phase filtering with an eight-point Hanning window.

2.3. Residual Selection

The residual codebook method described in the last section tries to represent an arbitrary residual by a linear combination of a limited number of prototype residuals. To better model the manifold characteristics of the residuals, the residual selection technique stores all residuals r_n seen in training into a table together with the corresponding feature vectors v_n that this time are composed of the line spectral frequencies and their deltas [4].

In operation phase, we have the current feature vector $\tilde{v}$ of the above described structure and choose one residual from the table by minimizing the square error between $\tilde{v}$ and all feature vectors seen in training ($S(v)$ is the sum over the squared elements of a vector v):

$$\tilde{r} = r_{\tilde{n}} \quad \text{with} \quad \tilde{n} = \underset{n\,=\,1,\ldots,N}{\arg\min}\, S(\tilde{v} - v_n) \; . \qquad (1)$$

Similar as discussed in Section 2.2, we also have to deal with an appropriate phase prediction in order to avoid signal discontinuities. At first, we train a Gaussian mixture model with I mixture components on the line spectral frequency representation of all target speaker spectra seen in the training corpus. Now, we introduce the vector $P(v) = [p(1|v), \ldots, p(I|v)]'$ which consist of the posterior probabilities that vector v belongs to the components $1, \ldots, I$, and a matrix $T = [T_1, \ldots, T_I]$ which contains the waveform templates T_1^I. In order to determine T, we minimize the following square error:

$$\varepsilon = \sum_{n=1}^{N} S(s(n) - TP(v_n)) \; ,$$

where $s(n)$ is the n^{th} speech frame normalized to a length of 10 ms.

During operation phase, for each frame, we are given a transformed feature vector $\tilde{v}$ and predict the corresponding waveform shape as $\tilde{s} = TP(\tilde{v})$. Finally, we substitute the phase spectrum of the current frame by the phase spectrum of $\tilde{s}$ and perform the smoothing described in the last section.

3. RESIDUAL PREDICTION: A NOVEL METHOD

Residual Selection and Smoothing

At the beginning of our work on the prediction of appropriate residuals, we wondered whether the line spectral frequencies and the corresponding residuals really are sufficiently correlated or if we should not expect a higher correlation between corresponding residuals of source and target speaker. Accordingly, in addition to the definition of the feature vector v as described in Section 2.3, we introduced the following normalized time-domain representation of the

source speech residuals to be used with the error criterion defined in Eq. 1:

$$v = \frac{|r - \bar{r}|}{\sqrt{S(r - \bar{r})}} ,$$

where r is the residual of a source speech frame and $\bar{r}$ its mean.

The observation that an inadequate phase treatment leads to a "rough" [6] or "harsh" [4] quality of the converted signal in spite of a sophisticated residual prediction let us emphasize the paragraphs dealing with phase prediction in Sections 2.2 and 2.3. Besides, the procedures described in these sections are only applied to voiced speech frames, whereas for unvoiced frames, one either copies the phases from the corresponding source frame [6] or selects unvoiced target speaker frames from the training corpus [4].

The novel technique described in this section is an integral approach that tries to simultaneously handle inaccuracies of the residual selection and phase prediction as well as the treatment of unvoiced frames by means of a time-variant residual smoothing.

We are given the sequence $\tilde{r}_1^K$ of predicted residual target vectors derived from Eq. 1, a sequence of scalars σ_1^K with $0 < \sigma_k \leq 1$ that are the voicedness degrees of the frames to be converted, and the voicedness gain α, cf. Section 4.2. At last, we obtain the final residuals by applying a normal distribution function to compute a weighted average over all residual vectors $\tilde{r}_1^K$, the deviation is defined by the product of voicedness degree and gain:

$$r_k^* = \sum_{\kappa=1}^{K} N(\kappa|k, \alpha\sigma_k) \cdot \tilde{r}_\kappa .$$

This equation can be interpreted as follows: In case of voiced frames ($\sigma \approx 1$), we obtain a wide bell curve that averages over several neighbored residuals, whereas for unvoiced frames ($\sigma \to 0$), the curve approaches a Dirac function, i.e., there is no local smoothing, the residuals and the corresponding phase spectra change chaotically over the time as expected in unvoiced regions.

4. EVALUATION

4.1. The Experimental Corpus

The corpus utilized in this work contains several hundred Spanish sentences uttered by a female and a male speaker. The speech signals were recorded in an acoustically isolated environment and sampled at a sample frequency of 16 kHz, cf. [9].

4.2. System Architecture and Parameter Settings

From the experimental corpus, we took 10 equivalent sentences of source and target speaker (about 42 s and 39 s) and extracted the pitch marks by applying the algorithm proposed in [10], since the described techniques require pitch-synchronous time frames. From the latter, we computed 16^{th} order line spectral frequency vectors (about 5k and 6k vectors, respectively). After aligning the corresponding sentences of the source and target speaker by means of dynamic time warping, we performed the parameter training according to [3] utilizing a Gaussian mixture model with 4 mixture components.

Since the training material in our experiment was not sufficient to reliably train the Gaussian mixture model of Section 2.2 with the settings specified in [6], we reduced the number of mixture components from 32 to 8. For the phase prediction of Section 2.3, we chose 16 mixture components. The voicedness degrees required in Section 3 were determined using the algorithm described in [11]; furthermore, we chose a voicedness gain of $\alpha = 3$.

Finally, the converted time frames computed by filtering the predicted residuals with the respective line spectral frequencies are concatenated using time domain pitch-synchronous overlap and add [12].

4.3. Subjective Evaluation

The goal of the subjective evaluation of the described residual prediction techniques is to answer two questions:

- Does the technique change the speaker identity in the intended way?
- How does a listener assess the overall sound quality of the converted speech?

The answers we want to find by means of the extended ABX test and the mean opinion score (MOS) evaluation described in [13].

From the experimental corpus, we took 3 sentences of the female and the male speaker (about 14 s and 13 s, respectively) and tried to convert these sentences in both directions: female-to-male (*f2m*) and male-to-female (*m2f*).

In our evaluation, we considered the following residual prediction techniques:

- *source residuals*: the source residuals are copied
- *target residuals*: time-aligned reference residuals are used
- *codebook method*
- *residual selection*
- *selection & smoothing*: residual selection based on line spectral frequencies and their deltas and smoothing
- *selection* & smoothing*: residual selection based on source residuals and smoothing

Now, 10 evaluation participants, 8 of whom specialists in speech processing, were asked if the converted voice sounds similar to the source or to the target voice or to neither of them (extended ABX test). Furthermore, they were asked to assess the overall sound quality of the converted speech on an MOS scale between 1 (very bad) and 5 (very good). Table 1 reports the results of the extended ABX test and Table 2 those of the MOS rating depending on the residual prediction technique and the gender combination.

%	source	target	neither
source residuals	20	10	70
reference residuals	0	79	21
codebook method	0	70	30
residual selection	0	70	30
selection & smoothing	0	85	15
selection* & smoothing	0	80	20

Table 1. Results of the extended ABX test

	m2f	f2m	total
source residuals	3.2	3.7	3.5
reference residuals	3.0	3.0	3.0
codebook method	1.6	1.9	1.8
residual selection	1.7	2.3	2.0
selection & smoothing	2.2	2.9	2.6
selection* & smoothing	2.2	2.8	2.5

Table 2. Results of the MOS test

5. INTERPRETATION

Addressing the first question that we asked in Section 4.3, we find that all assessed techniques succeed in converting the source voice to the target voice in more than 70% of the cases, cf. Table 1. The only exception is the use of the source residuals where the majority of the listeners had the impression of hearing a third speaker as we have already expected in Section 2.1. The novel residual selection technique with smoothing shows the highest conversion performance even higher than using the time-aligned reference residuals.

When we have a look at Table 2, we note that using unprocessed residuals produces the highest speech quality. Applying the reference residuals works slightly worse as the time-alignment based on dynamic time warping sometimes results in prosodic artifacts. However, we have to emphasize again that the reference speech is not given in a real world situation; we only considered this procedure to obtain a standard of comparison.

Having a look at the remaining techniques that succeeded in converting the source to the target voice, we see that the residual selection technique with smoothing outperforms the others in terms of speech quality although the absolute MOS scores (about 2.5) show that there is still a need for improvement. Past experiments on voice conversion techniques that explicitly are to "introduce as few distortions as possible" [14] while ignoring the success of the speaker identity conversion resulted in MOS scores about 3.0 [14], also on the same corpus [13].

Finally, we want to answer the question of Section 3 for the correlation between vocal tract parameters and residuals or source and target residuals, respectively. The outcomes of the ABX as well as the MOS evaluation show that there is no significant difference between the residual selection based on the one and the other. Consequently, both correlations are appropriate to our task, an observation that leads to the idea of using both the source residual and the converted vocal tract parameters to better predict the target residuals. This is to be subject to future investigations.

6. CONCLUSION

In this paper, we compared several residual prediction techniques to be used for voice conversion. The presented residual selection technique with smoothing outperforms the others in terms of voice conversion performance and speech quality. However, subjective tests show that, in general, voice conversion still perceptibly deteriorates the quality of the source speech whereas most of the compared techniques succeed in converting the speaker identity.

7. REFERENCES

[1] E. Moulines and Y. Sagisaka, "Voice Conversion: State of the Art and Perspectives," *Speech Communication*, vol. 16, no. 2, 1995.

[2] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker Adaptation for HMM-Based Speech Synthesis System Using MLLR," in *Proc. of the 3th ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.

[3] Y. Stylianou, O. Cappé, and E. Moulines, "Statistical Methods for Voice Quality Transformation," in *Proc. of the Eurospeech'95*, Madrid, Spain, 1995.

[4] H. Ye and S. J. Young, "High Quality Voice Morphing," in *Proc. of the ICASSP'04*, Montreal, Canada, 2004.

[5] L. M. Arslan and D. Talkin, "Voice Conversion by Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum," in *Proc. of the Eurospeech'97*, Rhodes, Greece, 1997.

[6] A. Kain and M. W. Macon, "Design and Evaluation of a Voice Conversion Algorithm Based on Spectral Envelope Mapping and Residual Prediction," in *Proc. of the ICASSP'01*, Salt Lake City, USA, 2001.

[7] A. Kain and M. W. Macon, "Spectral Voice Transformations for Text-to-Speech Synthesis," in *Proc. of the ICASSP'98*, Seattle, USA, 1998.

[8] H. Duxans, A. Bonafonte, A. Kain, and J. van Santen, "Including Dynamic and Phonetic Information in Voice Conversion Systems," in *Proc. of the ICSLP'04*, Jeju, South Korea, 2004.

[9] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge, "A First Step Towards Text-Independent Voice Conversion," in *Proc. of the ICSLP'04*, Jeju, South Korea, 2004.

[10] V. Goncharoff and P. Gries, "An Algorithm for Accurately Marking Pitch Pulses in Speech Signals," in *Proc. of the SIP'98*, Las Vegas, USA, 1998.

[11] "Adaptive Multi-Rate (AMR) Speech Transcoding," European Telecommunications Standards Institute, Sophia Antipolis, France, Tech. Rep. 3G TS 26.090, 1999.

[12] F. J. Charpentier and M. G. Stella, "Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation," in *Proc. of the ICASSP'86*, Tokyo, Japan, 1986.

[13] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge, "Time Domain Vocal Tract Length Normalization," in *Proc. of the ISSPIT'04*, Rome, Italy, 2004.

[14] M. Eichner, M. Wolff, and R. Hoffmann, "Voice Characteristics Conversion for TTS Using Reverse VTLN," in *Proc. of the ICASSP'04*, Montreal, Canada, 2004.