RESIDUAL PREDICTION

David Sündermann[1,2,3], *Harald Höge*[1], *Antonio Bonafonte*[2], *Helenca Duxans*[2]

[1]Siemens AG, Munich, Germany
[2]Universitat Politècnica de Catalunya, Barcelona, Spain
[3]University of Southern California, Los Angeles, USA
david@suendermann.com, harald.hoege@siemens.com, {antonio, hduxans}@gps.tsc.upc.edu

ABSTRACT

Residual prediction is a technique that aims at recovering the spectral details of speech that was encoded using parameterizations as linear predictive coefficients. Example applications of residual prediction are hidden Markov model-based speech synthesis or voice conversion. Our voice conversion experiments showed that only one of the seven compared techniques was capable of successfully converting the voice while achieving a fair speech quality (i.e. mean opinion score = 3).

1. INTRODUCTION

Several tasks of speech generation and manipulation as voice conversion [1] or hidden Markov model-based speech synthesis [2] are capable of dealing fairly well with speech that is encoded using parameter representations as mel frequency cepstral coefficients, linear predictive coefficients, or line spectral frequencies. These parameters aim at representing the vocal tract while the excitation is represented by the residual signal. Often, when generating (speech synthesis) or transforming (voice conversion) speech using one of the above parameterizations, the excitation is modeled very roughly using a single pulse in voiced regions and white noise with random phases in unvoiced regions. However, as this simple model may result in a synthetic sound of the voice and, furthermore, the residual seems to contain speaker-dependent information [3], it is reasonable to model the residual more carefully.

After dealing with the properties of our baseline voice conversion system in Section 2, we briefly describe seven residual prediction approaches in Section 3. Then, in Section 4, we examine these approaches using a Spanish cross-gender corpus by means of listening tests. The results of this evaluation are discussed in Section 5. It turns out that the technique based on unit selection outperforms the others in terms of voice conversion performance and sound quality.

This work has been partially funded by the European Union under the integrated project TC-Star - Technology and Corpora for Speech to Speech Translation - http://www.tc-star.org.

We would like to acknowledge the contribution of the numerous participants of the evaluation that is a part of this work.

2. THE BASELINE VOICE CONVERSION SYSTEM

2.1. Training Phase

A state-of-the-art technique based on linear transformation serves as our baseline system [4]. It requires parallel utterances of source and target speaker for training [5]. The speech data is split into pitch-synchronous frames using the pitch marking algorithm proposed in [6]. For extracting frames in unvoiced regions, this algorithm applies a linear interpolation between neighbored voiced regions. To improve the speech quality of the overlap and add technique used for the synthesis in Section 2.2, we always regard two successive pitch periods as being one frame as suggested in [3].

Now, the frame sequences of parallel source and target speaker utterances are aligned by means of dynamic time warping. Each frame is parameterized using linear predictive coefficients that are converted to line spectral frequencies that feature better interpolation properties.

Let x_1^M and y_1^M be parallel sequences of feature vectors of the source and target speech, respectively. Then, we use the combination of these sequences

$$z_1^M = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} x_M \\ y_M \end{pmatrix}$$

to estimate the parameters of a Gaussian mixture model $(\alpha_i, \mu_i, \Sigma_i)$ with I components for the joint density $p(x, y)$.

A list of the system parameters can be found in Table 1.

2.2. Conversion Phase

Here, we are given a source speaker's utterance that is processed as described in Section 2.1 yielding a sequence of feature vectors. Each source feature vector x is converted

parameter	description	value
f_s	sampling rate	16 kHz
$f_{0\mathrm{n}}$	norm fundamental frequency	100 Hz
q	quantization	16 bit
F	order of the line spectral frequencies	16
I	number of Gaussian mixtures for the linear transformation	4

Table 1. System parameters.

to a target vector y by the conversion function which minimizes the mean squared error between the converted source and the target vectors observed in training:

$$y = \sum_{i=1}^{I} p(i|x) \cdot (\mu_i^y + \Sigma_i^{yx} \Sigma_i^{xx\,-1}(x - \mu_i^x)) \,,$$

$$\text{where} \quad p(i|x) = \frac{\alpha_i N(x|\mu_i^x, \Sigma_i^{xx})}{\sum_{j=1}^{I} \alpha_j N(x|\mu_j^x, \Sigma_j^{xx})} \quad \text{and}$$

$$\Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix}; \; \mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix} .$$

The target line spectral frequency vectors are transformed to linear predictive coefficients that are used in the framework of linear prediction pitch-synchronous overlap and add [7] to generate the converted speech signal. Here, for each time frame, the respective underlying residual is required. In the following section, we want to deal with techniques that allow for predicting these residuals.

3. SEVEN TIMES RESIDUAL PREDICTION

3.1. The Trivial Solution: Copying Source 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. Macon, but they stated that "merely changing the spectrum is not sufficient for changing the speaker identity" [4]. Most of the listeners "had the impression that a 'third' speaker was created".

3.2. The Cheat: Copying Reference Residuals

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.

3.3. 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 [9].

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 with I_rc mixture components. 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:

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

During the conversion 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:

$$\tilde{m} = \sum_{i=1}^{I_\mathrm{rc}} \hat{m}_i p(i|\tilde{v}) \,.$$

3.4. Spectral Refinement

When applying the spectral refinement technique [10], at first, we train a Gaussian mixture model with I_sr mixture components on the line spectral frequency representation of all target speaker spectra seen in the training corpus. Now, we introduce the vector $P(v) = [p(1|v), \ldots, p(I_\mathrm{sr}|v)]'$ which consist of the posterior probabilities that vector v

belongs to the components $1, \ldots, I_{\mathrm{sr}}$, and a matrix $M = [M_1, \ldots, M_{I_{\mathrm{sr}}}]$ which consists of I_{sr} prototypes of logarithmic residual representations. In order to determine M, we minimize the following square error:

$$\varepsilon = \sum_{n=1}^{N} S(log(m_n) - MP(v_n)) \; ,$$

where $S(x)$ is the sum over the squared elements of a vector x, m_n is the residual magnitude spectrum of the n^{th} speech frame seen in training.

During the conversion phase, for each frame, we are given a transformed feature vector $\tilde{v}$ and predict the corresponding residual magnitude spectrum as

$$\tilde{m} = \exp(MP(\tilde{v})) \; .$$

3.5. Residual Selection

The residual codebook method and the spectral refinement described in the previous sections try to represent an arbitrary residual by a linear combination of a limited number of prototype residuals. Since both methods only deal with the residual magnitude spectrum, in addition, they have to apply a phase prediction; for details, please refer to the respective publications [9, 10].

To better model the manifold characteristics of the residuals and predict both magnitude as well as phase spectra at the same time, 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 [11].

In the conversion 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

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

3.6. Residual Smoothing

When listening to converted speech generated using the residual selection technique, in particular in voiced regions, we note a lot of artifacts and, sometimes, obviously improper residuals that occur due to the insufficient correlation between feature vectors and residuals. In voiced regions, the signal should be almost periodic; we do not expect the residuals to change abruptly. In unvoiced regions, as mentioned in Section 1, the residuals should feature a random phase spectrum that essentially changes from frame to frame. These considerations led to the idea of a voicing-dependent residual smoothing as proposed in [12].

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 voicing degrees of the frames to be converted, determined according to [13], and the voicing gain α. 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 standard deviation is defined by the product of gain and voicing degree:

$$r_k^* = \frac{\sum\limits_{\kappa=1}^{K} N(\kappa|k, \alpha\sigma_k) \cdot \tilde{r}_\kappa}{\sum\limits_{\kappa=1}^{K} N(\kappa|k, \alpha\sigma_k)} \; . \qquad (2)$$

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.

In order to be able to execute the summation, the vectors $\tilde{r}_1^K$ must have the same lengths. This is achieved by utilizing a normalization as suggested in [10], where all residuals are normalized to the norm fundamental frequency $f_{0\mathrm{n}}$, cf. Table 1.

3.7. Unit Selection

Although the residual smoothing approach essentially improves the speech quality (from a mean opinion score of 2.0 to 2.6, cf. Table 4) it is still insufficient for applications where the quality is of importance as for server-based speech synthesis with $f_s \geq 16$ kHz. Mainly, this is due to an oversmoothing caused when the voicing gain α is too large. This results in a deterioration of the articulation and increases the voicing of unvoiced sounds. However, when α is too small, the artifacts are not sufficiently suppressed, hence, the choice of α is based on a compromise.

Recently, we presented the unit selection approach [14], a technique that is well-known from concatenative speech synthesis [15]. Unit selection-based residual prediction is able to more reliably select residuals from the training database. Consequently, the selected residual sequence $\tilde{r}_1^K$ already contains less artifacts, thus, we can use smaller values for α and, hence, improve the quality of the converted speech.

Generally, in the unit selection framework, two cost functions are defined. In this case, the target cost $C^{\mathrm{t}}(r_k, \tilde{v}_k)$ is an estimate of the difference between the database residual r_k and that selected by means of feature vector $\tilde{v}_k$. The concatenation cost $C^{\mathrm{c}}(r_{k-1}, r_k)$ is an estimate of the quality of a join between the consecutive residuals r_{k-1} and r_k.

The searched residual sequence $\tilde{r}_1^K$ is determined by minimizing the sum of the target and concatenation costs applied to an arbitrarily selected sequence of K elements

	training		test	
	M	time	K	time
female	51,365	341.1 s	1,901	14.0 s
male	44,772	377.2 s	1,703	13.4 s

Table 2. Corpus statistics; M and K are the numbers of frames in the training and test data, respectively.

from the set of residuals seen in training, r_1^M, given the target feature sequence $\tilde{v}_1^K$:

$$\tilde{r}_1^K = \arg\min_{r_1^K} \sum_{k=1}^{K} C^{\mathrm{t}}(r_k, \tilde{v}_k) + C^{\mathrm{c}}(r_{k-1}, r_k)$$

4. EVALUATION

4.1. Experimental Corpus

The corpus utilized in this work contains 100 Spanish sentences uttered by a female and a male speaker, cf. [16]. It was designed to provide baseline voices for Spanish speech synthesis, e.g. in the UPC text-to-speech system [17]. For the corpus statistics, cf. Table 2.

4.2. 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?

We want to find the answers by means of the extended ABX test and the mean opinion score (MOS) evaluation described in [18].

We performed both, female-to-male (*f2m*) and male-to-female (*m2f*) voice conversion using the corpus described in Section 4.1. Now, 27 evaluation participants, 25 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). In doing so, they were asked to ignore the recording conditions, the sound or synthesis quality of the samples, the speaking style, and the prosody. Furthermore, they assessed the overall quality of the converted speech on an MOS scale between 1 (bad) and 5 (excellent).[1]

Table 3 reports the results of the extended ABX test and Table 4 those of the MOS rating depending on the residual prediction technique and the gender combination.

[1]The subjective evaluation was carried out using the web interface `http://suendermann.com/unit`.

[%]	source	target	neither
source residuals	20	10	70
reference residuals	0	79	21
residual codebook	0	70	30
spectral refinement	0	**83**	17
residual selection	0	70	30
residual smoothing	0	**85**	15
unit selection	2	**83**	15

Table 3. 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
residual codebook	1.6	1.9	1.8
spectral refinement	2.0	2.0	2.0
residual selection	1.7	2.3	2.0
residual smoothing	2.2	2.9	2.6
unit selection	2.8	3.2	**3.0**

Table 4. Results of the MOS test

5. INTERPRETATION

Addressing the first question that we asked in Section 4.2, 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 3. 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 3.1. Spectral refinement, residual smoothing, and unit selection showed almost the same conversion performance: In about 85% of the cases, the target voice was recognized which is even higher than using the time-aligned reference residuals.

When we have a look at Table 4, we note that using unprocessed residuals produces the highest speech quality. Applying the reference residuals works 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 unit selection technique outperforms the others in terms of speech quality and achieves a quality similar to that of using the reference residuals (MOS = 3.0). Past experiments on voice conversion techniques that explicitly are to "introduce as few distortions as possible" [19] while ignoring the success of the speaker identity conversion resulted in MOS

scores about 3.0 [19], also on the same corpus [18]. Consequently, the achieved speech quality is state-of-the-art.

6. CONCLUSION

In this paper, we compared seven residual prediction techniques and applied them to voice conversion. We found that the unit selection-based approach outperforms the others in terms of speech quality. Furthermore, this technique resulted in a voice conversion performance that was one of the best achieved in the evaluation.

7. REFERENCES

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

[2] K. Tokuda, H. Zen, and A. W. Black, "An HMM-Based Speech Synthesis System Applied to English," in *Proc. of the IEEE Speech Synthesis Workshop*, Santa Monica, USA, 2002.

[3] A. Kain, *High Resolution Voice Transformation*, Ph.D. thesis, Oregon Health and Science University, Portland, USA, 2001.

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

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

[6] 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.

[7] E. Moulines and F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones," *Speech Communication*, vol. 9, no. 5, 1990.

[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 Island, South Korea, 2004.

[9] 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.

[10] H. Ye and S. J. Young, "Quality-Enhanced Voice Morphing Using Maximum Likelihood Transformations," *To appear in IEEE Trans. on Speech and Audio Processing*, 2005.

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

[12] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge, "A Study on Residual Prediction Techniques for Voice Conversion," in *Proc. of the ICASSP'05*, Philadelphia, USA, 2005.

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

[14] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, and A. W. Black, "Residual Prediction Based on Unit Selection," in *Proc. of the ASRU'05*, Cancun, Mexico, 2005.

[15] A. J. Hunt and A. W. Black, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database," in *Proc. of the ICASSP'96*, Atlanta, USA, 1996.

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

[17] A. Bonafonte, I. Esquerra, A. Febrer, J. A. R. Fonollosa, and F. Vallverdú, "The UPC Text-to-Speech System for Spanish and Catalan," in *Proc. of the ICSLP'98*, Sydney, Australia, 1998.

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

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