RESIDUAL PREDICTION BASED ON UNIT SELECTION

David Sündermann[1,2], *Harald Höge*[1], *Antonio Bonafonte*[2], *Hermann Ney*[3], *Alan W Black*[4]

[1]Siemens Corporate Technology, Munich, Germany
[2]Universitat Politècnica de Catalunya, Barcelona, Spain
[3]RWTH Aachen – University of Technology, Aachen, Germany
[4]Carnegie Mellon University, Pittsburgh, USA

`david@suendermann.com, harald.hoege@siemens.com, antonio.bonafonte@upc.edu,`
`ney@cs.rwth-aachen.de, awb@cs.cmu.edu`

ABSTRACT

Recently, we presented a study on residual prediction techniques that can be applied to voice conversion based on linear transformation or hidden Markov model-based speech synthesis. Our voice conversion experiments showed that none of the six compared techniques was capable of successfully converting the voice while achieving a fair speech quality. In this paper, we suggest a novel residual prediction technique based on unit selection that outperforms the others in terms of speech quality (mean opinion score = 3) while keeping the conversion performance.

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.

In a recent study [4], we compared six residual prediction techniques by means of a subjective test investigating their voice conversion performance and the quality of the converted speech. We concluded that

- The best speech quality (mean opinion score [MOS] = 3.5) is achieved when simply using the residuals of the source speech rather than predicting target residuals. However, as stated above, the residuals contain an important part of the speaker identity, consequently, only in 10 % of the cases, the test subjects recognized the target voice, whereas in 70 % of the cases, they had the impression of hearing a third speaker's voice, cf. Table 1.

- The techniques that predicted the target residuals succeeded in converting the voice but resulted in an essential loss of speech quality. Residual selection and smoothing showed the highest voice conversion performance (in 85 % of the cases, the target voice was recognized) and speech quality (MOS = 2.6), cf. Table 1.

Consequently, the most important goal was to achieve a better speech quality without affecting the conversion performance.

After describing the voice conversion baseline system in Section 2, we analyze the shortcomings of the residual selection and smoothing technique and extend it by applying the unit selection paradigm in Sections 3 to 5. Finally, in Section 6, we compare voice conversion performance and speech quality of the residual selection and smoothing technique with those of the unit selection technique by means of a subjective evaluation. We show that the new approach based on unit selection outperforms the other in terms of speech quality (MOS = 3) while keeping the conversion performance.

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.

	extended ABX [%]			MOS
	source	target	neither	
source residuals	20	10	70	3.5
residual selection	0	70	30	2.0
selection & smoothing	0	85	15	2.6

Table 1. Extract from the results of the study on residual prediction techniques. For the terms *extended ABX* and *MOS*, cf. Section 6.2.

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 [5]. It requires parallel utterances of source and target speaker for training [6]. The speech data is split into pitch-synchronous frames using the pitch marking algorithm proposed in [7]. 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 2.

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 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:

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

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 2. System parameters.

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

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

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 [8] 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. RESIDUAL SELECTION

Preassumption of most of the residual prediction techniques is a considerable correlation between the feature vectors and the corresponding residuals[1].

The residual selection technique [9] stores all residuals r_m seen in training into a table together with the corresponding feature vectors v_m. The latter are composed of the line spectral frequencies and their deltas.

In the conversion phase, we are given the current feature vector $\tilde{v}$ of the above described structure and choose the corresponding 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):

$$\tilde{r} = r_{\tilde{m}} \quad \text{with} \quad \tilde{m} = \underset{m\,=\,1,\ldots,M}{\arg\min} \, S(\tilde{v} - v_m) \,. \qquad (1)$$

4. RESIDUAL SMOOTHING

When listening to converted speech generated using the residual selection technique, in particular in voiced regions,

[1] The linear predictive coding technique is based on a source-filter model that tries to separate the effect of excitation and vocal tract; ideally, both should be uncorrelated. Hence, the success of residual prediction techniques based on this correlation indicates a shortcoming of the model.

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 [4].

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 [10], 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 ($\sigma \rightarrow 0$), 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 [11], where all residuals are normalized to the norm fundamental frequency $f_{0\mathrm{n}}$, cf. Table 2.

5. UNIT SELECTION

Although the residual smoothing approach essentially improves the voice conversion performance (from 70 % to 85 % of the cases the target speaker was recognized) and the speech quality (from a mean opinion score of 2.0 to 2.6, cf. Table 1), the latter 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[2]. However, when α is too small, the artifacts are not sufficiently suppressed, hence, the choice of α is based on a compromise. For the corpus and the settings described in Sections 2 and 6.1, we found that $\alpha = 3$ is a reasonable choice.

When there would be a possibility to more reliably select residuals from the training database so that the selected residual sequence $\tilde{r}_1^K$ already contains less artifacts, we could use smaller values for α and, hence, improve the quality of the converted speech. Certainly, the most appropriate residual sequence is one that we have seen in training and that fulfills the optimization criterion of Eq. 1 at the same time. Of course, this will only apply if the converted feature sequence $\tilde{v}_1^K$ is identical to a feature sequence seen in training. Since this will never be the case, we should weaken these conditions and allow the residual sequence to be composed of several subsequences seen in training whose endings fit to each other. Furthermore, we should also take suboptima of Eq. 1 into account in order to obtain subsequences of reasonable lengths. This approach is called *unit selection*, a technique that is widely used in concatenative speech synthesis [12].

Generally, in the unit selection framework, two cost functions are defined. The target cost $C^{\mathrm{t}}(u_k, t_k)$ is an estimate of the difference between the database unit u_k and the target t_k which it is supposed to represent. The concatenation cost $C^{\mathrm{c}}(u_{k-1}, u_k)$ is an estimate of the quality of a join between the consecutive units u_{k-1} and u_k.

In speech synthesis, the considered units are phones, syllables, or even whole phrases, whereas in the residual prediction case, we set our base unit length to be a single speech frame, since this allows for being independent of additional linguistic information about the processed speech as for instance the phonetic segmentation. Furthermore, the cost functions can simply be defined by interpreting the residuals as database units, i.e. $u := r$. In the following sections, we describe the properties of the cost functions used for unit selection-based residual prediction and how we come to the final residual sequence.

5.1. The Target Cost Function

Similar to the residual selection procedure described in Section 3, the appropriateness of a residual r seen in training for being selected is determined based on the distance between the corresponding feature vector v, and that one which represents the properties of the converted frame, $\tilde{v}$. Furthermore, we also want to take fundamental frequency and energy of the considered residual into account. This is to minimize the extent of the signal processing to produce the prosodic characteristics of the converted speech and, thus, avoid distortions of the natural waveform. According to [12], the target cost C^{t} is calculated as the weighted sum of the distances between the considered features of the target and the candidate:

[2] This sounds like a paradox; but when you take into account that the voicing degree σ is continuous and always greater than 0, the multiplication with a large voicing gain α in Eq. 2 may transform an unvoiced into a voiced sound.

$$C^{\mathrm{t}}(u,t) := C^{\mathrm{t}}(r,(\tilde{v},\tilde{f}_0,\tilde{S})) = \qquad (3)$$

$$= w_1 d(v(r),\tilde{v}) + w_2 d(f_0(r),\tilde{f}_0) + w_3 d(S(r),\tilde{S}) \, .$$

$\tilde{f}_0$ and $\tilde{S}$ are the target fundamental frequency and energy that, in the case of residual prediction for voice conversion, can be derived from the respective parameters of the source speech frame by applying individual conversion rules. For instance, a simple conversion rule for the fundamental frequency is the multiplication with the ratio between the mean fundamental frequencies of target and source that were determined in training.

For the weights holds (cf. Eq. 5)

$$w_1 + w_2 + w_3 \leq 1; \quad w_1, w_2, w_3 \geq 0 \, .$$

This makes sure that the sum of the cost functions' weights including that of the concatenation cost described in Section 5.2 is always 1.

$v(r)$, $f_0(r)$, and $S(r)$ are the feature vector that corresponds to the candidate residual according to the table generated in Section 3 and fundamental frequency and energy of the residual.

d is the Mahalanobis distance that compensates for differences of range and amount of variation between the features used in Eq. 3:

$$d(x,y) = \sqrt{(x-y)'\Sigma^{-1}(x-y)} \, . \qquad (4)$$

Σ is the covariance matrix computed using the respective features of all residuals seen in training.

5.2. The Concatenation Cost Function

The cost for concatenating the residuals r_{k-1} and r_k is defined using the residual normalization introduced in [4]:

$$C^{\mathrm{c}}(r_{k-1},r_k) = (1-w_1-w_2-w_3)\cdot S\{n(r_k)-n(r_{k-1})\}$$

$$\text{with} \quad n(r) = \frac{|r-\bar{r}|}{\sqrt{S(r-\bar{r})}} \, . \qquad (5)$$

When r_{k-1} and r_k are residuals that belonged to successive frames in the training data, the concatenation should be optimal, hence, in this special case, we define $C^{\mathrm{c}}(r_{k-1},r_k)$ to be 0. Again, since both considered residuals must have the same number of samples, a normalization to the norm fundamental frequency $f_{0\mathrm{n}}$ is carried out, cf. Section 4.

$\bar{r}$ is a vector whose elements equal the mean value of r's 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 3. Corpus statistics; M and K are the numbers of frames in the training and test data, respectively.

5.3. Finding the Optimal Residual Sequence

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 from the set of residuals seen in training, r_1^M, given the target feature sequences $\tilde{v}_1^K$, $\tilde{f}_{0\,1}^{\,K}$, and $\tilde{S}_1^K$:

$$\tilde{r}_1^K = \arg\min_{r_1^K} \sum_{k=1}^{K} C^{\mathrm{t}}(r_k,(\tilde{v_k},\tilde{f}_{0\,k},\tilde{S}_k)) + C^{\mathrm{c}}(r_{k-1},r_k) \, . \qquad (6)$$

5.4. On the Computational Complexity

Due to the special structure of the function to be optimized that is a sum whose addends are only dependent on the parameters of the current position k and the preceding position $k-1$, this optimization problem can be solved by means of dynamic programming. Although this prevents us from the intractable treatment of M^K paths, it turns out that we still face a high computational effort: The full solution of Eq. 6 requires

$$O \approx K \cdot M^2 \cdot (8F^2 + 4F + 6\frac{f_{\mathrm{s}}}{f_{0\mathrm{n}}}) \text{ ops} \, . \qquad (7)$$

For a description of the parameters in this formula and the respective values from the experimental corpus described in Section 6.1, see Tables 2 and 3. In the following example, we want to look at female-to-male conversion, i.e., K belongs to the female speaker and M to the male in Table 3.

When we use these example parameter settings and run the computation on a computer that executes $3\ \mathrm{Gops/s}$, it would take *one and a half months,* i.e. 280,000 times real-time. Consequently, we must find ways to simplify the algorithm in order to build a real-time system:

- By introducing a pruning that only considers the 5 best hypotheses, we are able to reduce the real-time factor (RTF) to 31.
- Instead of utilizing the Mahalanobis distance (Eq. 4), we apply the Euclidean distance (RTF $= 11$).
- We restrict the calculation of the concatenation cost in Eq. 5 to the first of the two signal periods contained in the processed residuals, cf. Section 2.1 (RTF $= 5.8$).

- The norm fundamental frequency $f_{0\mathrm{n}}$ used to transform the residuals to vectors of identical lengths, cf. Sections 4 and 5.2, can be duplicated without noticeably affect the behavior of the concatenation cost function ($\mathrm{RTF} = 3.8$).

- Taking into account that almost half of the durations of the speech signal to be converted and that used for training is actually non-speech (silence or noise), we obtain a real-time factor of 0.8.

Hence, at least on very fast computers, the algorithm is real-time-able.

5.5. Residual Smoothing

As predicted at the beginning of this section where we motivated the introduction of the unit selection paradigm for residual prediction, informal listening tests showed that the output of the unit selection features essentially less artifacts than that of the residual selection discussed in Section 3. However, since there are still audible signal discontinuities, the application of the residual smoothing described in Section 4 is still recommendable. It turns out, that the smoothing gain α can effectively be decreased due to the already smoother input residual sequence. We determined $\alpha = 1.5$ for the unit selection approach (as opposed to $\alpha = 3.0$ for the residual selection), thus, the output features a higher naturaleness and better articulatory properties. In the following section, this statement is to be confirmed by means of a subjective evaluation.

6. EVALUATION

6.1. Experimental Corpus

The corpus utilized in this work contains 100 Spanish sentences uttered by a female and a male speaker, cf. [13]. It was designed to provide baseline voices for Spanish speech synthesis, e.g. in the UPC text-to-speech system [14]. For the corpus statistics, cf. Table 3.

6.2. Subjective Evaluation

By means of a subjective evaluation, we want to compare the best residual prediction technique of the study we presented in [4], the selection and smoothing technique, with the novel unit selection approach. The goal of the subjective evaluation 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?

[%]	source	target	neither
selection & smoothing	0	85	15
unit selection & smoothing	2	83	15

Table 4. Results of the extended ABX test

	m2f	f2m	total
selection & smoothing	2.2	2.9	2.6
unit selection & smoothing	2.8	3.2	3.0

Table 5. Results of the MOS test

We want to find the answers by means of the extended ABX test described in [15] and an MOS test [16].

We performed both, female-to-male (*f2m*) and male-to-female (*m2f*) voice conversion using the corpus described in Section 6.1. 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).[3]

Table 4 reports the results of the extended ABX test and Table 5 those of the MOS rating depending on the residual prediction technique and the gender combination.

6.3. Interpretation

The outcomes of the extended ABX test show almost no differences between both techniques: In about 85 % of the cases, the source voice was successfully converted to the target voice. Having a look at the speech quality results, we note that the novel unit selection approach clearly outperforms the residual selection and smoothing: In particular, for the harder task, the male-to-female conversion, we obtained a considerable gain of 0.6 MOS points. The overall result ($\mathrm{MOS} = 3.0$, i.e. a *fair* speech quality) is still essentially worse than natural speech ($\mathrm{MOS} = 4.8$), but already suitable for many applications that do not require high fidelity speech, for instance in telecommunications.

7. CONCLUSION

In this paper, we introduced a novel technique for residual prediction based on unit selection. We applied this technique to cross-gender voice conversion and showed that, in

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

terms of speech quality, the new approach clearly outperforms the residual selection and smoothing method which was the best-performing out of six residual prediction techniques according to a recently published study. The conversion performance remained unaffected.

8. 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] 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.

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

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

[7] 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.

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

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

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

[11] H. Ye and S. J. Young, "Transformation-Based Voice Morphing," *Submitted to IEEE Trans. on Speech and Audio Processing*, 2005.

[12] 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.

[13] 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.

[14] 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.

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

[16] "Methods for Subjective Determination of Transmission Quality," Tech. Rep. ITU-T Recommendation P.800, ITU, Geneva, Switzerland, 1996.