TEXT-INDEPENDENT VOICE CONVERSION BASED ON UNIT SELECTION

David Sündermann[1,2,3], *Harald Höge*[1], *Antonio Bonafonte*[2], *Hermann Ney*[4], *Alan Black*[5], *Shri Narayanan*[3]

[1]Siemens Corporate Technology, Munich, Germany
[2]Universitat Politécnica de Catalunya, Barcelona, Spain
[3]University of Southern California, Los Angeles, USA
[4]RWTH Aachen – University of Technology, Aachen, Germany
[5]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 shri@sipi.usc.edu

ABSTRACT

So far, most of the voice conversion training procedures are text-dependent, i.e., they are based on parallel training utterances of source and target speaker. Since several applications (e.g. speech-to-speech translation or dubbing) require text-independent training, over the last two years, training techniques that use non-parallel data were proposed. In this paper, we present a new approach that applies unit selection to find corresponding time frames in source and target speech. By means of a subjective experiment it is shown that this technique achieves the same performance as the conventional text-dependent training.

1. INTRODUCTION

Voice conversion is the adaptation of the characteristics of a source speaker's voice to those of a target speaker [1]. Conventional voice conversion techniques are text-dependent. I.e., they need equivalent utterances of source and target speaker as training material which can be automatically aligned by dynamic time warping [2]. This procedure is necessary since the training algorithms require corresponding time frames for feature extraction.

The precondition of having equivalent utterances is inconvenient and often results in expensive manual work: New speech material must be recorded or bilingual speakers are required. E.g. when applying voice conversion to speech-to-speech translation [3], we want the target voice that is synthesized by a text-to-speech system to be identical to the source speaker's voice. Since source and target language are different, it is very unlikely to have parallel utterances of both speakers. Here, the usage of text-independent training is inevitable.

In this paper, we review former approaches to text-independent voice conversion emphasizing advantages and shortcomings of the respective techniques, cf. Section 3. We then derive the new technique based on unit selection in Section 4. Finally, text-dependent and unit-selection based text-independent training are compared using a subjective test in Section 5.

2. VOICE CONVERSION BASED ON LINEAR TRANSFORMATION

The most popular voice conversion technique is the application of a linear transformation to the spectra of pitch-synchronous speech frames [2]. 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. Mostly, these spectral features are mel frequency cepstral coefficients [2, 4] or line spectral frequencies [5, 6]. According to the linear predictive source-filter model [7], the features are supposed to represent the vocal tract contribution, i.e. mainly the phonetic content of the processed speech. In addition, the speaker-dependence of the excitation contribution has to be taken into account. This leads to the issue of *residual prediction*, for details see [8].

In order to reliably train the Gaussian mixture model parameters, parallel sequences of training feature vectors are required. The parallelity is necessary because otherwise there is no guaranty that in the conversion phase the phonetic contents remain unchanged – as a result, the speech would become inarticulate or even unintelligible. Parallel feature vectors can be derived from parallel utterances by applying dynamic time warping [9].

3. TEXT-INDEPENDENT VOICE CONVERSION: STATE-OF-THE-ART

3.1. Automatic Segmentation and Mapping of Artificial Phonetic Classes

Now, we consider text-independent voice conversion, i.e., we are given source and target speech based on non-parallel utterances, and the goal is to find frames that phonetically correspond to each other. If we had a phonetic segmentation of the respective speech, we could take a set of frames from corresponding segments of source and target speech as input of a conventional parameter training. However, we do not have phonetically segmented speech, thus, we must find a way to automatically produce an appropriate segmentation.

In our previous work on this basic concept [10], the k-means algorithm was applied to length-normalized magnitude spectra of pitch-synchronous speech frames resulting in a segmentation into artificial phonetic classes. Figure 1 displays the speech waveform of the word *Arizona*. In this example, the clustering algorithm was to distribute the speech frames among eight classes. It automatically assigned class 0 to silence, class 1 to the phoneme /i/, class 2 to /e/, etc.

This automatic segmentation is executed for both source and target speech resulting in K source classes and L target classes.

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.

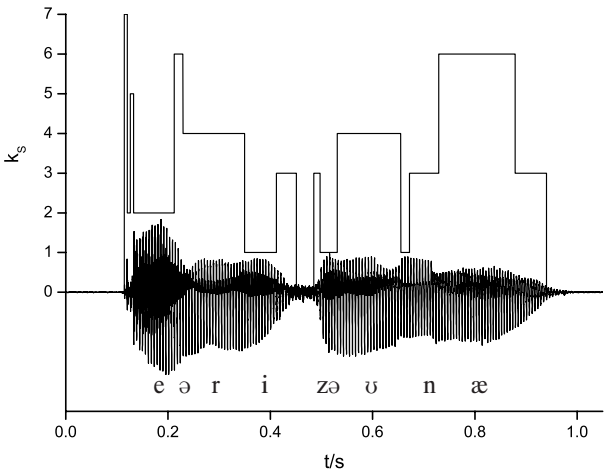

Fig. 1. Automatic class segmentation of the word *Arizona*

As discussed above, now the objective is to find corresponding, i.e. phonetically equivalent source and target classes. This is done by comparing the centroids of the source and target clusters, i.e. the prototype magnitude spectra of each source and target class, respectively. For each target cluster $l \in \{1, \ldots, L\}$, we find one corresponding source cluster:

$$k(l) = \underset{\kappa\,=\,1,\ldots,K}{\arg\min}\, S(\bar{X}_\kappa - \bar{Y}_l); \quad S(X) = \sqrt{X'X}\,. \qquad (1)$$

As we suggested in [11], aside from a sole cluster mapping, we must produce full parallel frame sequences to be used to reliably train the linear transformation parameters. This is done by shifting each target cluster in such a way that its centroid $\bar{Y}$ coincides with the corresponding source centroid $\bar{X}$. Finally, for each shifted target cluster member $Y' = Y - \bar{Y} + \bar{X}$, we determine the nearest member of the mapped source class, X, in terms of the Euclidean distance. The desired spectrum pairs consist of the respective unshifted target spectra Y and the determined corresponding source spectra X:

$$X = \arg\min_{\chi} S(\chi - Y - \bar{X} + \bar{Y})\,.$$

3.2. Text-independent voice conversion using a speech recognizer

The fully automatic segmentation and mapping does not require any phonetic knowledge about the considered languages. This fact is advantageous on the one hand since the technique can be applied to arbitrary languages without additional data. On the other hand, more information about the phonetic structure of the processed speech could lead to a more reliable mapping between source and target frames.

Consequently, Ye and Young [12] proposed to use a speaker-independent hidden Markov model-based speech recognizer to label each frame with a state index such that each source or target speaker utterance is represented by a state index sequence. If the text of these utterances is known, this can be done by means of forced alignment resulting in more reliable state sequences.

In a second step, subsequences are extracted from the set of target sequences to match the given source state index sequences using a simple selection algorithm. This algorithm favors longer matching sequences to ensure a continuous spectral evolution of the selected target speech frames. The latter are derived from the state indices considering the frame–state mapping delivered by the speech recognizer. Based on these parallel frame sequences, the conventional linear transformation parameter training is applied.

4. TEXT-INDEPENDENT VOICE CONVERSION BASED ON UNIT SELECTION

So far, the algorithms for parallelizing frame sequences of source and target speech tried to

- extract the phonetic structure underlying the processed speech,
- find a mapping between the phonetic classes of source and target speech,
- and transform the class mapping back to a frame mapping.

All these steps may produce errors that accumulate and may lead to a poor parameter estimation or even to a convergence failure of the parameter estimation algorithm as reported in [11]. Hence, it would be helpful to avoid the detour through the class layer and, instead, find the mapping only using frame-based features. These features could be those used for the spectral representation of speech frames mentioned in Section 2: mel frequency cepstral coefficients or line spectral frequencies.

Given the source speech feature sequence x_1^M, we would simply be able to determine the best-fitting sequence of corresponding target feature vectors $\tilde{y}_1^M$ by selecting from an arbitrary target feature sequence y_1^N. This could be done similarly to the class mapping in Equation 1:

$$\tilde{y}_m = \underset{n\,=\,1,\ldots,N}{\arg\min}\, S(x_m - y_n); \quad m = 1, \ldots, M\,.$$

However, as we learned in Section 3.2, in this selection one should also take the continuity of the spectral evolution into account.

To find the best possible compromise between minimizing the distance of source and target features and achieving a maximal continuity, we make use of the well-studied *unit selection* framework widely used for concatenative speech synthesis [13] and recently applied to residual prediction for text-dependent voice conversion [14].

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

In speech synthesis, the considered units are phones, syllables, or even whole phrases, whereas in the frame alignment task, 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. Hence, the cost functions can be defined by interpreting the feature vectors as database units, i.e. $t := x$ and $u := y$.

For determining the target vector sequence $\tilde{y}_1^M$ best-fitting the source sequence x_1^M, we have to minimize the sum of the target and concatenation costs applied to an arbitrarily selected sequence of vectors taken from the non-parallel target sequence y_1^N.

Furthermore, since all compared units are of the same structure (they are feature vectors) and dimensionality, the cost functions may be represented by Euclidean distances, thus, we finally have

$$\tilde{y}_1^M = \arg\min_{y_1^M} \sum_{m=1}^{M} \left\{ \alpha S(y_m - x_m) + (1-\alpha) S(y_{m-1} - y_m) \right\}.$$

Here, the parameter α is for adjusting the tradeoff between fitting accuracy of source and target sequence and the spectral continuity criterion.

5. EXPERIMENTS

We already mentioned that one of the main applications of text-independent voice conversion is the post-processing step of a speech-to-speech translation system as for instance in the European project TC-Star [3]. This project involves the world's three most spoken languages: English, Mandarin and Spanish. In the scope of three evaluation campaigns, the single modules' performances are to be assessed and compared among the numerous project partners as well as external participants.

In the recently performed first campaign, the text-dependent voice conversion algorithm described in Section 2 was assessed with respect to voice conversion performance and sound quality [15]. Here, two different residual prediction techniques were compared. For the following investigations, we decided to focus on the technique that achieved a higher speech quality: residual transformation based on vocal tract length normalization.

5.1. Residual Transformation Based on Vocal Tract Length Normalization

Unlike the ideal source-filter model that assumes the voice source to be a white noise, some investigations have shown that producing an appropriate excitation is crucial for generating natural voice-converted speech [5, 8]. Unfortunately, it is not enough to directly use the source speech frame's residual and use it as excitation of the transformed (linear predictive) feature vector because it contains an essential amount of speaker-dependent information. Consequently, efforts have been undertaken to predict suitable target residuals based on the target feature vectors.

In this work, we applied vocal tract length normalization (VTLN) [16], a technique that is widely used for speaker normalization in speech recognition. Although VTLN has already been used for voice conversion [10], the novelty of the recent investigation was its application to the residuals instead of to the speech frames.

The successful application of VTLN to the residuals of speech may seem unmotivated for the following reasons:

- As the name implies, vocal tract length normalization is to change the length, or more generally, the shape of the *vocal tract*. According to the source-filter model, the vocal tract is represented by the features rather than by the residuals. Consequently, an application to the residuals should hardly change the speech characteristics.

- Furthermore, according to experiences in speech recognition, applying VTLN in conjunction with a linear transformation in feature space should not help since the linear transformation already compensates for the effect of speaker-dependent vocal tract lengths and shapes [17].

Surprisingly, perceptive tests have shown that the application of VTLN to the residuals before filtering with the vocal tract features strongly influences the voice identity and may influence voice properties as gender or age.

Unlike the residual prediction techniques discussed in [8], the contribution of the VTLN-based residual transformation to the speech quality deterioration is almost negligible compared to those of the linear transformation and the pitch and time-scale modification. However, due to the small number of parameters used for the VTLN (in general only two: *warping factor* and *fundamental frequency ratio*, cf. [18]), for particular source / target voice combinations the residuals are not reliably transformed.

5.2. Subjective Experiments

In this section, we want to compare text-dependent voice conversion with the unit selection-based text-independent conversion regarding the voice conversion performance and speech quality.

For this purpose, we utilized the same database as used for the aforementioned TC-Star evaluation campaign: A Spanish speech corpus consisting of 50 phonetically balanced utterances of four speakers (two male and two female). From each of the 50 sentences, 10 sentences were randomly selected to be used as test data; the remaining data was used for training. For the text-dependent technique, we took exactly the same speech samples as for the TC-Star evaluation to have a standard of comparison.

For the text-independent case, we randomly split the training data into 20 different sentences for the source and for the target speaker, respectively, to have a real-world scenario where the source and target texts are completely different.

From the possible twelve source / target speaker combinations, we selected the same four pairs as in the TC-Star evaluation (each gender combination once).

To carry out the comparison, we performed a subjective test according to [19] where 13 subjects (exclusively speech processing specialists) participated.

The subjects were asked to rate the presented speech samples regarding two aspects:

- For each conversion method and gender combination, the subjects listened to speech sample pairs from the converted and the target voice and were to rate their similarity on a five-point scale (1 for *different* to 5 for *identical*). This single score is referred to as $d_{c,t}$. The same was done for the respective source / target speaker pairs resulting in $d_{s,t}$. According to the TC-Star specification [19], both scores are combined to take the similarity of the involved voices before the conversion into account and obtain a better estimation of the conversion's contribution to the final voices' similarity:

$$D_{c,t} = \begin{cases} 1.0 & : \quad d_{c,t} < d_{s,t} \\ \emptyset & : \quad d_{c,t} = d_{s,t} = 5 \\ \dfrac{5 - d_{c,t}}{5 - d_{s,t}} & : \quad \text{otherwise} \end{cases}$$

Finally, this score is averaged over all subjects resulting in the voice conversion score (VCS) which corresponds to a normalized distance or error measure ($0 \leq \text{VCS} \leq 1$).

- To assess the overall speech quality, we used the mean opinion score (MOS) well-known from telecommunications [20]. For each speech sample, the subjects are asked to rate the speech quality on a five-point scale (1 for *bad*, 2 for *poor*, 3 for *fair*, 4 for *good*, 5 for *excellent*). The average over all participants is the MOS.

VCS	text-dependent	text-independent	TC-Star
m2m	**0.51**	**0.24**	**0.53**
m2f	0.67	0.60	0.85
f2m	0.73	0.81	0.88
f2f	0.85	0.84	0.91
Σ	0.69	0.62	0.79

Table 1. Results of the subjective test: voice conversion performance (VCS $= 0$ is success, VCS $= 1$ is failure)

MOS	text-dependent	text-independent	TC-Star
m2m	2.4	2.3	3.4
m2f	2.9	2.7	3.3
f2m	2.7	2.5	3.2
f2f	2.8	2.7	2.9
Σ	2.7	2.6	3.2
clean	4.7		4.6

Table 2. Results of the subjective test: overall speech quality

In Table 1 and 2, the results of the subjective tests are displayed for the four gender combinations and the two compared techniques. As a standard of comparison, the results of the TC-Star evaluation based on the same text-dependent technique are shown.

5.3. Interpretation

Having a look at the voice conversion scores displayed in Table 1, we note that the outcomes highly depend on the particular voice combination. Interestingly, the scores consistently become worse the higher the female contribution is. This statement agrees with former investigations [11] and encourages a stronger emphasis on female voice conversion in future investigations. Surprisingly, in this test, the text-independent voice conversion outperforms the text-dependent conversion. Furthermore, we note a large gap between the identical test sets of the TC-Star evaluation and the text-dependent case. This gap is also obvious when focusing on the overall speech quality results shown in Table 2. In average, the TC-Star and text-dependent scores differ by 0.5 MOS points which could be interpreted as a considerable deterioration of the speech quality. However, as the speech samples in both tests were identical, the reason must be due to the evaluation framework. The major differences between the two subjective tests are

- the subjects' scientific background (naïve subjects in the case of TC-Star vs. speech processing experts in the other case) and
- the fact that in the TC-Star evaluation one more technique with considerably different characteristics was assessed. This technique featured a better voice conversion performance but worse speech quality. Hence, the participants tended to divide the samples into 3 classes: bad, fair, and good speech quality, where *fair* was associated with the VTLN-based residual transformation and *good* with the natural (clean) speech.

In contrast to the latter, in the present case, we only assessed two kinds of speech, converted[1] and natural. Now, the subjects tended to divide the samples into two opposite categories: *bad* for converted and *good* for natural speech.

[1]According to Table 2, text-dependent and text-independent conversion were hardly distinguishable in terms of speech quality.

6. CONCLUSION

In this paper, we presented a novel approach for text-independent parameter training for linear transformation-based voice conversion. This approach uses the unit selection framework well-studied in speech synthesis. A subjective test has shown that the novel text-independent approach achieves the same performance as the conventional text-dependent training based on dynamic time warping.

7. REFERENCES

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

[2] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 2, 1998.

[3] H. Höge, "Project Proposal TC-STAR - Make Speech to Speech Translation Real," in *Proc. of the LREC'02*, Las Palmas, Spain, 2002.

[4] T. Toda, A. W. Black, and K. Tokuda, "Spectral Conversion Based on Maximum Likelihood Estimation Considering Global Variance of Converted Parameter," in *Proc. of the ICASSP'05*, Philadelphia, USA, 2005.

[5] 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.

[6] 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.

[7] A. Acero, "Source-Filter Models for Time-Scale Pitch-Scale Modification of Speech," in *Proc. of the ICASSP'98*, Seattle, USA, 1998.

[8] 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.

[9] C. S. Myers and L. R. Rabiner, "A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected Word Recognition," *Bell System Technical Journal*, vol. 60, no. 7, 1981.

[10] D. Sündermann, H. Ney, and H. Höge, "VTLN-Based Cross-Language Voice Conversion," in *Proc. of the ASRU'03*, Virgin Islands, USA, 2003.

[11] 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.

[12] H. Ye and S. J. Young, "Voice Conversion for Unknown Speakers," in *Proc. of the ICSLP'04*, Jeju Island, South Korea, 2004.

[13] 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.

[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. Bonafonte, H. Höge, I. Kiss, A. Moreno, U. Ziegenhain, H. v. d. Heuvel, H.-U. Hain, and X. S. Wang, "TTS: Specifications of LR, Systems' Evaluation and Modules Protocols," in *Submitted to the LREC'06*, Genoa, Italy, 2006.

[16] D. Pye and P. C. Woodland, "Experiments in Speaker Normalization and Adaptation for Large Vocabulary Speech Recognition," in *Proc. of the ICASSP'97*, Munich, Germany, 1997.

[17] M. Pitz and H. Ney, "Vocal Tract Normalization Equals Linear Transformation in Cepstral Space," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, 2005.

[18] D. Sündermann, G. Strecha, A. Bonafonte, H. Höge, and H. Ney, "Evaluation of VTLN-Based Voice Conversion for Embedded Speech Synthesis," in *Proc. of the Interspeech'05*, Lisbon, Portugal, 2005.

[19] A. Bonafonte, H. Höge, H. S. Tropf, A. Moreno, H. v. d. Heuvel, D. Sündermann, U. Ziegenhain, J. Pérez, and I. Kiss, "TC-Star: Specifications of Language Resources for Speech Synthesis," Tech. Rep., 2005.

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