Text-Independent Cross-Language Voice Conversion

David Sündermann[1,2], *Harald Höge*[1], *Antonio Bonafonte*[2], *Hermann Ney*[3], *Julia Hirschberg*[4]

[1]Siemens Corporate Technology, Munich, Germany
[2]Technical University of Catalonia, Barcelona, Spain
[3]RWTH Aachen, Aachen, Germany
[4]Columbia University, New York, USA

david@suendermann.com harald.hoege@siemens.com
antonio.bonafonte@upc.edu ney@cs.rwth-aachen.de julia@cs.columbia.edu

Abstract

So far, cross-language voice conversion requires at least one bilingual speaker and parallel speech data to perform the training. This paper shows how these obstacles can be overcome by means of a recently presented text-independent training method based on unit selection. The new method is evaluated in the framework of the European speech-to-speech translation project TC-Star and achieves a performance similar to that of text-dependent intra-lingual voice conversion.

Index Terms: voice conversion, unit selection, TC-Star.

1. Introduction

Voice conversion is the transformation of a source speaker's voice to that of a target speaker. Usually, the conversion is performed in two steps:

- Model parameters are trained based on training speech data of source and target speaker,
- these parameters specify the characteristics of a conversion function that is applied to source speech data and aims at transforming the latter to sound similar to the target voice.

As suggested one decade ago by Stylianou et al. [1], we use a conversion function based on a linear transformation in feature space. The parameters of the conversion function are derived using a joint Gaussian mixture model (GMM) of source and target speech features. This approach is still state-of-the-art and regarded as robust and capable of producing high speech quality [2].

As features, line spectral frequencies (LSFs) have shown to have superior properties compared to other features commonly used in speech processing (as mel frequency cepstral coefficients or linear predictive coefficients) [2]. Furthermore, most voice conversion systems apply pitch-synchronous processing, since this allows for using standard pitch modification techniques to change prosodical properties of the source speaker to become closer to those of the target speaker. I.e., a speech frame (which is basis for computing a feature vector) consists of one pitch period[1].

When training parameters of a joint GMM, for each source feature vector of the training data we need a corresponding target feature vector. The conventional solution of this problem strongly limits the applicability of the voice conversion technology:

So far, most training procedures use parallel training utterances of source and target speaker, align the speech by means of dynamic time warping (DTW) and, finally, derive feature vector sequences whose contents are treated as being parallel. We call this approach *text-dependent* [3].

However, several applications require a training based on arbitrary utterances of source and target speaker (*text-independent* approach). In particular, this is necessary when source and target speaker use different languages which is referred to as *cross-language* voice conversion [4].

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`.

[1]In our system, we take *two* pitch periods as a frame which supports more robust pitch modification.

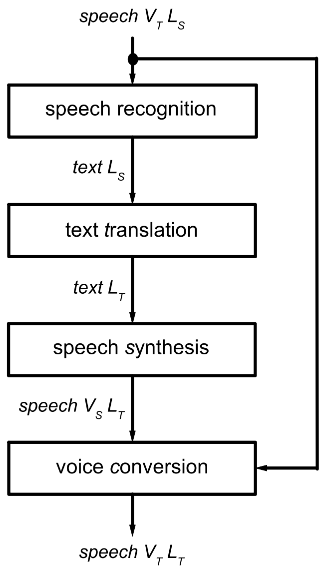

Figure 1: The components of a speech-to-speech translation system with voice conversion. V and L stand for voice and language, S and T for source and target.

The aim of the European speech-to-speech translation project TC-Star [5] is to recognize the speech of an English-speaking person, translate it to a target language (Spanish or Mandarin) and then convert it to speech using a text-to-speech synthesizer. Finally, the standard voice of the synthesizer is to be converted to the voice of the source speaker to preserve its individuality. Here, source and target are based on different languages. Hence, we face the cross-language voice conversion task. Figure 1 shows the components of a speech-to-speech translation system with voice conversion.

2. Related Work

Interestingly, the very first investigations on cross-language voice conversion in the beginning of the nineties also focused on the speech-to-speech translation task [4][2]. At that time, ATR – where the authors of the latter paper were working – was developing a so-called *interpreting telephone*. This was the name of a speech-to-speech translation system applied to telephone conversations and integrated a cross-language voice conversion module to preserve speaker recognizability across languages.

[2]This paper's authors also seem to be the first dealing with voice conversion in general, see [8].

authors	technique	training type (flags)	speech quality (MOS)	conversion performance (ABX)
Abe et al. [4]	codebook mapping	text-independent (T_2L_2)	?	fair*
Mashimo et al. [6]	linear transformation	text-dependent (T_0L_1)	fair*	good*
Sündermann et al. [7]	VTLN	text-independent (T_2L_2)	fair	fair

Table 1: Former cross-language voice conversion techniques. The type flags (see Table 2) were introduced to precisely characterize the underlying training procedure that is important for the target application of voice conversion, cf. Section 1.
*according to the corresponding intra-lingual system

This first attempt was based on a codebook mapping that used a discrete representation of the acoustic feature space. To the best of our knowledge, there were no investigations carried out dealing with the technique's speech quality. Besides, it was not sufficiently shown whether this approach is able to successfully convert voice characteristics. The results reported were based on objective error measures that are not standardized and sometimes hardly correlate with the perceptive similarity, cf. [3]. Subjective experiments using the described codebook mapping technique reported successful gender transformation from male to female and 61% successfully transformed examples for male-to-male conversion using an ABX test[3] [8].
In Table 1, the discussed cross-language voice conversion techniques are compared in terms of training type, speech quality and conversion performance.
More than a decade later, Japanese researchers (some of them also at ATR) continued the investigations on cross-language voice conversion and applied the linear transformation-based conversion function introduced in Section 1. However, unlike their predecessors, Mashimo at al. [6] avoided the text independence problem by using bilingual (Japanese/English) speakers as source speakers. The conversion function was trained on parallel Japanese utterances of source and target speaker and applied to English source speech in conversion phase. The only difference to text-dependent intra-lingual voice conversion are the distinct phoneme sets of source and target language. The corresponding intra-lingual baseline system described in [10] achieved a fair speech quality (mean opinion score 2.9) and a conversion performance of about 90% on an ABX scale.
In 2003, we investigated the application of vocal tract length normalization (VTLN), a technique which is widely used in speech recognition, to cross-language voice conversion [7]. Due to the very small number of conversion parameters (2 to some dozens), a phonetic clustering algorithm could be applied that led to a mapping of speech segments in non-parallel speech. Consequently, the proposed algorithm was text-independent. On the other hand, the limited parameter number only allowed for converting the main voice characteristics (as gender and age) and, hence, sometimes did not properly convert voices. An ABX test showed that about 50% of the cases were successfully transformed. The speech quality of VTLN-based voice conversion was found to be fair; subjective listening tests reported a mean opinion score of 3.0 [11].

3. Cross-Language Voice Conversion Training Based on Unit Selection

3.1. Motivation

The goal of the following investigation is to find a way to change the training interface of a state-of-the-art text-dependent and intra-lingual voice conversion system to become text-independent and applicable to the cross-language task. The resulting system should combine the training type of Abe et al. with the performance of Mashimo et al. as required in applications like speech-to-speech translation.

3.2. The Concept

In a recent study on text-independent voice conversion parameter training [12], we presented a technique with unit selection which achieves almost the same performance on the intra-lingual task as text-dependent training based on DTW. It takes two sequences of feature (LSF) vectors representing source and target speech, x_1^M and y_1^N, and selects from the latter the feature vector sequence $\tilde{y}_1^M$ that optimally corresponds to the source sequence. This is done by taking two criteria into account:

- The distance between source and corresponding target features (*target cost*) is minimum (optimal correspondence).
- The distance to the neighbors of the corresponding target feature vector (*concatenation cost*) is minimum (optimal naturalness).

Mostly, these optima do not coincide, and we must get by with a compromise between both: We search for the minimum of the weighted sum of target and concatenation cost for each source feature vector:

$$\tilde{y}_1^M = \underset{y_1^M}{\arg\min} \sum_{m=1}^{M} \Big\{ \alpha S(y_m - x_m) + (1-\alpha) S(y_{m-1} - y_m) \Big\}. \quad (1)$$

Here, $S(w)$ is the Euclidean distance

$$S(w) = \sqrt{w'w} \quad (2)$$

and $0 \leq \alpha \leq 1$ is a weight influencing the trade-off between target and concatenation cost.
The second aforementioned criterion is supposed to select naturally smooth segments[4] from the target feature vector sequence y_1^M. Since the optimal concatenation we expect is that of vectors which are neighbored in the original target speech, y_m and y_{m+1}, we regard the concatenation cost of such a vector pair to be zero rather than to be the Euclidean distance according to Eq. 2.
On the other hand, the Euclidean distance between two identical vectors is zero, a fact that would support repititions of the same vectors. To avoid this effect that could lead to undesirable voicing of the respective signal section, the concatenation cost between identical vectors is assigned infinity.
After determining $\tilde{y}_1^M$, conventional voice conversion parameter training is performed as discussed in Section 1.
Unlike text-dependent training based on bilingual speakers (Section 2), this time, the joint GMM is already cross-lingual, consequently, there is no language-dependent mismatch between training and conversion.

3.3. Time Behavior

Although the computational characteristics have not played a role in this study so far, we would like to mention that the discussed text-independent training turns out to be very time-consuming.

[3] Kain and Macon [9] showed that the ABX test often is not powerful enough to assess the performance of a voice conversion technique. Therefore, in the framework of TC-Star, a mean opinion score is used as described in Section 4.3.

[4] or *units*; that is, where the term *unit selection* stems from. This paradigm is well-known from concatenative speech synthesis where optimal speech units are selected and concatenated, cf. [13].

text flags

T_0	text-dependent: parallel training utterances
T_1	semi-text-independent: parallel training utterances but treated as non-parallel
T_2	text-independent: non-parallel training utterances

language flags

L_0	training and test in the same language
L_1	training in one language and test in another language (requires bilingual speakers)
L_2	source and target voice use different languages in training

Table 2: Type flags for voice conversion training

The structure of Eq. 1 allows for applying dynamic programming that makes the problem tractable. However, unlike well-known applications of dynamic programming, e.g. for DTW, in the case of the unit selection-based training, the search space is considerably larger: Conventionally, in the former case, the possible successors of a feature vector y_n are limited to the set $\{y_n, y_{n+1}, y_{n+2}\}$, whereas in the latter case all vectors are allowed. This leads to a time complexity of $\mathbf{O}(M \cdot N^2)$.
Let us consider an example: For the male-to-female conversion described in Section 4, we used about 400s speech data of source and target speaker, respectively. Taking the different fundamental frequencies of both speakers into account (or rather: their different frame lengths), we had $M = 5.7 \cdot 10^4$ and $N = 8.4 \cdot 10^4$ vectors. According to the aforementioned complexity, we had to compute about $4 \cdot 10^{14}$ times the expression in the curly braces of Eq. 1. After perfoming several steps to reduce the complexity[5], the computation still took more than 80 hours on a 3GHz Intel Xeon processor corresponding to a real time factor of about 730.

4. Evaluation

In this section, we discuss the evaluation of the presented technique in the framework of the periodical TC-Star evaluation campaigns. The first campaign took place in September 2005, the second is currently being performed (March/April 2006).
The evaluation is carried out by the independent research institute ELDA and concerns all components of the speech-to-speech translation system introduced in Section 1: speech recognition, machine translation, speech synthesis, and voice conversion – which will be focus of the following considerations.

4.1. TC-Star Evaluation Campaign I

This evaluation campaign was limited to intra-lingual voice conversion; details can be found in [12]. Since test corpus as well as the metrics for assessing the voice conversion performance were altered in the second campaign, the first evaluation's results are not suitable as baseline.

4.2. TC-Star Evaluation Campaign II

It the second campaign, we participated with both intra-lingual as well as cross-language voice conversion (for the characteristics refer to Table 3).
As already mentioned in Section 1, we applied pitch-synchronous signal processing and used LSFs as spectral features. Pitch marks were determined using the algorithm of Goncharoff and Gries [14]. According to [15], in addition to the feature conversion that is carried out by the linear transformation described in Section 1, we have to consider the speaker dependence of the underlying residual. Just applying the converted features to the unchanged source

[5] by exchanging x_1^M and y_1^N and pruning

	intra-lingual	**cross-language**
training type (flag)	text-dependent (T_0)	text-independent (T_2)
conversion type (flag)	intra-lingual (L_0)	cross-language (L_2)
source language	English	Spanish
target language	English	English
alignment technique	DTW	unit selection
sampling rate	16kHz / 16bit	
speakers	2 female, 2 male (bilingual professionals)	
training data amount	$\approx 400\text{s}$ per speaker and language	
pitch mark extraction		
- of training data	automatic, supervised	
- of test data	automatic, manually corrected	
features	LSF	
order	32	
GMM mixtures	4	
covariance type	diagonal	
residual conversion	VTLN	

Table 3: Characteristics of the voice conversion techniques assessed in the second TC-Star evaluation campaign. For the training type flags, see Table 2.

residuals might lead to a voice that is different from both source and target speaker. The aforementioned publication studies several techniques that successfully change the speaker identity, however, all these techniques considerably deteriorate the speech quality.
Since from our point of view the signal quality was of higher priority, we decided to apply VTLN (introduced in Section 2) to the residuals of the source speech, a technique that often is able to essentially contribute to change the source speaker identity towards that of the target speaker while barely affecting the speech quality. In conjunction with the linear transformation, we expected a reasonable conversion performance [12].
The following steps were to further enhance the system's performance:
Pitch tracking. Correct and consistent pitch marks are crucial for a good synthesis based on time domain pitch-synchronous overlap and add, which is the synthesis technique our voice conversion is based on [16]. Furthermore, already in the training phase, pitch mark errors can lead to a poor estimation of the GMM parameters. Therefore, after automatically determining the pitch marks by means of the algorithm mentioned in Section 4.2, we selected only those training speech files for the parameter training whose pitch marks had been reliably determined. For the test speech files, where correct pitch marks are much more important, they were manually corrected.
Voicing information. According to Ye and Young [2], most of the speaker-dependent information is carried by the voiced signal parts, whereas the unvoiced parts are almost speaker-independent. Consequently, it makes sense to copy the source speech signal in unvoiced parts and only apply the conversion to voiced sections. In order to take the potential (but sparse) speaker dependence of unvoiced sounds into account, we applied VTLN also to unvoiced sounds.
Feature dimensionality. The aforementioned issue of residual prediction only emerged because we describe the spectral envelope of a speech frame by means of a low-dimensional feature vector. The error we make by considering this envelope being the original signal is the residual which carries the original signal's spectral details and phase information. The better the feature vector represents the original speech frame, the smaller the residual's contribution becomes. This is done by increasing the dimensionality of the feature vector. However, the higher the vector's dimensionality is, the more unreliable are the GMM parameters trained on the

	MOS_Q	MOS_S
intra-lingual	3.3	2.4
cross-language	3.5	2.0
source voice	4.7	1.6

Table 4: Results of the second TC-Star evaluation campaign: overall speech quality and conversion performance

vector sequences. Due to the relatively large amount of training data available, it was possible to use a vector dimensionality (LSF order) of 32 without perceptibly affecting the speech quality.

4.3. Evaluation Metrics

The second TC-Star evaluation campaign for voice conversion was based on the following subjective error measures:

- To assess the overall speech quality, we used the mean opinion score (MOS) [17]. For each speech sample, the subjects were 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 samples and participants is referred to as MOS_Q.
- To evaluate the conversion performance, 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*). The average over all samples and participants is the mean opinion score MOS_S.

4.4. Corpus

The voice conversion corpus consists of recordings of four professional bilingual speakers (two female and two male). They uttered about 200 Spanish and 160 British English phrases (about $900\mathrm{s}$ and $800\mathrm{s}$ of speech) that were recorded using a high-quality distant microphone, a close-talk microphone and a Laryngograph at $96\mathrm{kHz}$, $24\mathrm{bit}$ sampling rate (for the experiments, a down-sampled version was used, cf. Table 3). From this corpus, 10 utterances were selected for testing, the remaining data served for training. For intra-lingual voice conversion, the training data was based on the English recordings only, for cross-language voice conversion, the source speaker data was English, that of the target Spanish. For details of the evaluation procedure refer to [18].

4.5. Results

In Table 4, we compare intra-lingual with cross-language voice conversion in terms of speech quality and conversion performance. The results are based on the opinion of 9 subjects whose mother tongue is British English[6]. As standard of comparison, we also give results of the unconverted (source) voice. The latter features the highest achievable speech quality but, at the same time, is a lower bound for the similarity to the target.

5. Interpretation

5.1. Speech Quality

The speech quality of both intra-lingual and cross-language voice conversion is between good and fair; the TC-Star goal of at least $\mathrm{MOS}_Q = 3$ was fulfilled [19].

5.2. Intra-Lingual vs. Cross-Language

Interestingly, the speech quality of text-independent cross-language voice conversion outperformed that of the text-dependent intra-lingual type. On the other hand, intra-lingual achieved a higher similarity score than cross-language voice conversion. Both effects might be attributed to the nature of the text-independent training method:

The cost minimization described in Eq. 1 encourages low target costs, i.e. low distances between source and corresponding target vector. The more training data is available, the smaller become these distances. For an infinite amount of training data, we expect them to tend to zero[7]. However, the more similar corresponding source and target vectors are, the less speaker-dependent information can be trained from them. For the limit case, where we have equivalent source and target vectors, we get zero vectors and identity matrices as parameters of the linear transformation. In this case, the converted feature vectors were equivalent to the source vectors, i.e., we would produce the source speech as output.

This consideration suggests to carefully select amount and nature of the training data for the text-independent training method to make sure that as much as possible speaker-dependent information can be learned from the data.

6. References

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

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

[3] 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.

[4] M. Abe, K. Shikano, and H. Kuwabara, "Cross-Language Voice Conversion," in *Proc. of the ICASSP'90*, Albuquerque, USA, 1990.

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

[6] M. Mashimo, T. Toda, K. Shikano, and N. Campbell, "Evaluation of Cross-Language Voice Conversion Based on GMM and STRAIGHT," in *Proc. of the Eurospeech'01*, Aalborg, Denmark, 2001.

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

[8] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice Conversion through Vector Quantization," in *Proc. of the ICASSP'88*, New York, USA, 1988.

[9] A. Kain and M. 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] T. Toda, J. Lu, H. Saruwatari, and K. Shikano, "Straight-Based Voice Conversion Algorithm Based on Gaussian Mixture Model," in *Proc. of the ICSLP'00*, Beijing, China, 2000.

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

[12] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-Independent Voice Conversion Based on Unit Selection," in *Proc. of the ICASSP'06*, Toulouse, France, 2006.

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

[14] 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.

[15] 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.

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

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

[18] A. Bonafonte, H. Höge, H. 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.

[19] A. Bonafonte, H. Höge, I. Kiss, A. Moreno, D. Sündermann, U. Ziegenhain, J. Adell, P. Agüero, H. Duxans, D. Erro, J. Nurminen, J. Pérez, G. Strecha, M. Umbert, and X. Wang, "TC-STAR: TTS Progress Report," Tech. Rep., 2005.

[6] The evaluation's final results will be based on 20 subjects.

[7] or at least a considerable small limit. This effect is to be investigated in a future study.