TC-Star: Cross-Language Voice Conversion Revisited

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

In the framework of the European speech-to-speech translation project TC-Star, one of the research tasks is cross-language voice conversion. In the recent second evaluation campaign, five participants presented their intra-lingual as well as cross-language voice conversion systems that were applied to three languages. In this paper, we discuss the results of Siemens' submissions and describe the underlying system characteristics.

1. Introduction

The aim of the European speech-to-speech translation project TC-Star (Höge, 2002) is to recognize speech of an English-speaking person, translate it to a target language (Spanish or Mandarin) and then transform it to speech using a text-to-speech synthesizer. Finally, the baseline voice of the synthesizer is to be converted to the voice of the source speaker to preserve its individuality. The latter process is referred to as *voice conversion*. Figure 1 shows the components of a speech-to-speech translation system with voice conversion.

According to the source-filter model (Acero, 1998), most voice conversion techniques are based on a vocal tract and excitation model of the source and target speaker or its differences, respectively. The model parameters are estimated in a training step where speech frames of source and target speaker with equivalent phonetic contents are required (Stylianou et al., 1995). To achieve this phonetic equivalence, in general, utterances of both speakers based on the same text (parallel utterances) are used and then aligned using dynamic time warping (DTW) or, if the text is known, forced alignment. Since this approach, which we call *text-dependent* (Sündermann et al., 2004), is only applicable if both speakers speak the same language, it is mostly used for intra-lingual voice conversion.

Naturally, when dealing with speech-to-speech translation, source and target speaker do not speak the same language, so we face the cross-language task. However, cross-language voice conversion does not necessarily imply a need for text-independence as Mashimo et al. (2001) have suggested:

Their Japanese-English voice conversion algorithm was trained using bilingual source speakers. In training, parallel English utterances of source and target speaker were used, whereas in conversion phase, the source speaker spoke Japanese. According to our experience, speech quality and conversion success are often independent of if training and conversion are applied to the same language or not. Similar results were shown in another publication of the aforementioned authors (Mashimo et al., 2002).

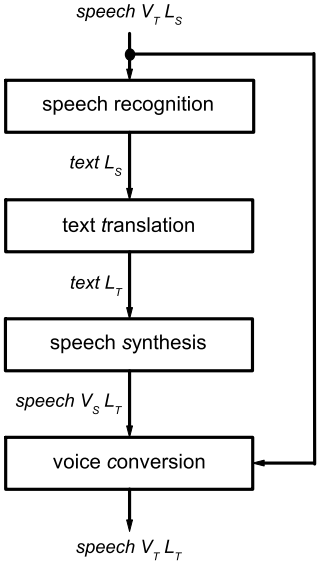

Figure 1: The components of a speech-to-speech translation system with voice conversion. V and L stand for voice and language, S and T for source and target.

Often, there is no bilingual source speaker available, in particular if we deal with uncommon languages as, e.g., Croatian (Black et al., 2002), Arabic (Waibel et al., 2003) or Pashto (Franco et al., 2003). Furthermore, in a speech-to-speech translation framework, the source speaker, whose voice is used for building a text-to-speech synthesizer, is a professional, carefully selected according to various subtle criteria, cf. (Bonafonte et al., 2005b). This is one of the reasons why we also investigated text-independent solutions for voice conversion training.

In the following section, we compare Siemens' intralingual

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`.

and cross-language training methods which were applied in the second TC-Star evaluation campaign. Evaluation results and their interpretation are discussed in Section 3.

2. From Intra-Lingual to Text-Independent Cross-Language Voice Conversion

2.1. Text-Dependent Intra-Lingual Voice Conversion

As suggested one decade ago by Stylianou et al. (1995), we use a conversion function based on 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 (Ye and Young, 2004).
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) (Ye and Young, 2004).
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 come 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.
As already mentioned in Section 1., our intra-lingual voice conversion is based on the text-dependent paradigm, i.e., we use parallel utterances of source and target speaker which are aligned by means of DTW, derive parallel feature vector sequences and, finally, train the parameters of the joint GMM.

Vocal Tract Length Normalization. According to (Sündermann et al., 2005), in addition to the feature conversion by means of linear transformation, we have to consider the speaker dependence of the underlying residual. Just applying the converted features to the unchanged source 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 vocal tract length normalization (VTLN) 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 (Sündermann et al., 2003). In conjunction with the linear transformation, we expected a reasonable conversion performance (Sündermann et al., 2006a).
VTLN is well-known from speech recognition where it serves as speaker normalization technique assuming that an important part of the speaker dissimilarity is caused by differences in vocal tract lengths. A normalization of the latter can be achieved by warping the frequency axis of the magnitude spectrum. We applied a linear warping function, whose slope, the warping factor, can be estimated by minimizing the spectral distance between warped source and target speech.
Although studies from the speech recognition domain (Pitz and Ney, 2005) have shown that using both linear transformation and VTLN cannot help because the latter is included in the former, we recently found out that this does not apply to VTLN transforming the residual spectrum rather than the spectral envelope which is represented by the features (Sündermann et al., 2006c).

Voicing information. According to Ye and Young (2004), 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 the spectra of unvoiced sounds using a separate warping factor.

2.2. Text-Independent Cross-Language Voice Conversion

We already mentioned in Section 1. that the main difference between text-dependent intra-lingual and text-independent cross-language voice conversion is the missing parallelism between training utterances of source and target speaker. Hence, for our cross-language system, we took the same architecture as for the intra-lingual one but applied the recently presented text-independent voice conversion parameter training with unit selection (Sündermann et al., 2006a). 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:

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

Here, S is the Euclidean distance

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

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

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

Table 1: Characteristics of Siemens' voice conversion techniques assessed in the scope of the second TC-Star evaluation campaign.

naturally smooth segments[1] 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 repetitions 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 1.), this time, the joint GMM is already cross-lingual, consequently, there is no language-dependent mismatch between training and conversion.

3. Evaluation

In this section, we discuss the evaluation of the above described techniques in the framework of the second TC-Star evaluation campaign performed in March and April 2006.

The evaluation was carried out by the independent research institute ELDA and concerned all components of the speech-to-speech translation system introduced in Section 1.: speech recognition, machine translation, speech synthesis, and voice conversion.

The voice conversion evaluation concerned all three TC-Star-related languages, British English, Spanish, Mandarin; five research groups – Chinese Academy of Sciences, IBM, Nokia, Siemens (ourselves), and the Technical University of Catalonia – took part.

Siemens participated with both text-dependent intra-lingual (English[2] and Spanish) as well as text-independent cross-language voice conversion (Spanish to English), for the characteristics refer to Table 1.

3.1. Evaluation Metrics

The evaluation was based on the following subjective error measures:

- To assess the overall speech quality, we used the mean opinion score (MOS) well-known from telecommunications (itu, 1996). 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.

3.2. Corpus

The voice conversion corpus consisted of recordings of four professional bilingual speakers (two female and two male). They uttered about 200 Spanish and 160 British English phrases (about 900s and 800s of speech) that were recorded using a high-quality distant microphone, a close-talk microphone and a Laryngograph at 96kHz, 24bit sampling rate

[1] 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. (Hunt and Black, 1996).

[2] This paper's focus is the English evaluation, since in this language, we submitted results of intra-lingual as well as cross-language voice conversion. Furthermore, English achieved the highest number of submissions (nine) from four competitors.

	MOS_Q	MOS_S
text-dependent intra-lingual	3.1	2.4
text-independent cross-language	3.4	2.0

Table 2: Results of the second TC-Star evaluation campaign: overall speech quality and conversion performance

(for the experiments, a down-sampled version was used, cf. Table 1). 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 (Bonafonte et al., 2005b).

3.3. Results

In Table 2, we compare text-dependent intra-lingual with text-independent cross-language voice conversion in terms of speech quality and conversion performance. The results are based on the opinion of 14 subjects whose mother tongue is British English.

To simplify interpreting these outcomes, in Figure 2, the performance of all competing systems is displayed as points in an MOS_S-MOS_Q coordinate system. In addition to our system's results (**intra-lingual** and **cross-language**), the following points are denoted to serve as standards of comparison:

- **source**. This is the source speech that naturally achieved the highest speech quality but, at the same time, the lowest similarity to the target.

- **IBM**. Only one more group, IBM, submitted a text-independent system. It was based on a method similar to the VTLN-based technique described in Section 2.1. but did not make use of a linear transformation in spectral feature domain as we did; for details, refer to (Chazan et al., 2006). This submission achieved the highest speech quality but the lowest similarity score.

- **synthesis**. The speech processing group at the Technical University of Catalonia built a text-to-speech synthesis system exclusively based on the speech corpus described in Section 3.2. As the amount of available speech data was very limited compared to conventional speech synthesis corpora that usually use several hours of data, the achieved sound quality was relatively poor. On the other hand, it did not convert source speech to sound like the target but directly took speech segments (units) from the given target speech and concatenated them. Therefore, the similarity score was very high.

- ***optimum***. This is the region where an optimal voice conversion system is located.

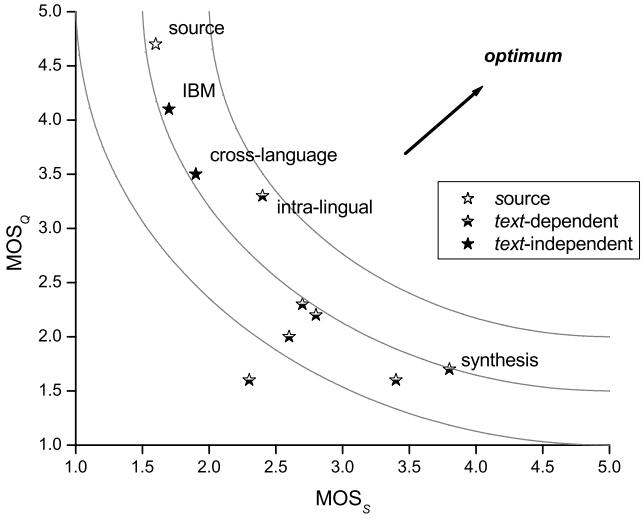

Figure 2: Results of the second TC-Star evaluation campaign. The gray lines are the set of points with distance $d = \{3, 3.5, 4\}$ from the optimum $\text{MOS}_Q = \text{MOS}_S = 5$.

4. Interpretation

4.1. Speech Quality

The speech quality of both Siemens submissions, intra-lingual and cross-language voice conversion, is between good and fair; the TC-Star goal of at least $\text{MOS}_Q = 3$ was fulfilled (Bonafonte et al., 2005a).

4.2. Text-Dependence vs. Text-Independence

As we already discussed in (Sündermann et al., 2006b), it is surprising that the speech quality of text-independent cross-language voice conversion outperformed that of the text-dependent intra-lingual type. For the similarity score, the outcomes were the other way around. We attributed both effects 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 expected them to tend to zero. 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.

In order to investigate the validity of the aforementioned assumption that the distances between corresponding feature vectors tend to zero for an infinite amount of training data, we conducted the following objective experiment:

For different amounts of training data (t_{tr} between 9 and 700 s) taken from the corpus described in Section 3.2., we measured the average Euclidean distance between corresponding feature vectors derived by means of the technique

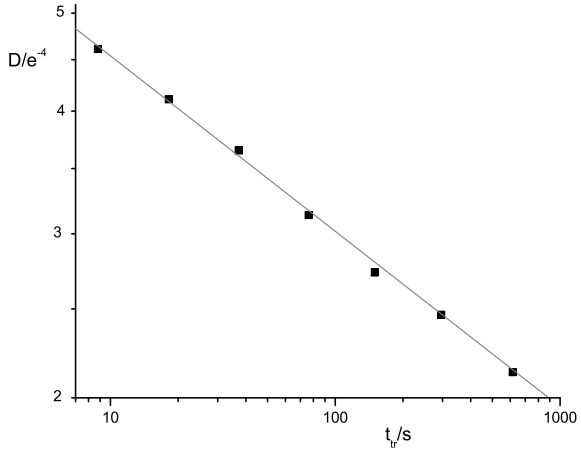

Figure 3: Text-independent parameter training: average distance between corresponding source and target feature vectors depending on the amount of training data.

from Section 2.2. (D) and obtained the result displayed in Figure 3.

We observe that in double logarithmic representation, the measuring points are almost located on a straight line which means that the average feature vector distance depends on the amount of training data as

$$D(t_{\mathrm{tr}}) = a t_{\mathrm{tr}}^{-b} \quad \text{with} \quad a, b > 0\,. \tag{3}$$

Since we have

$$\lim_{t_{\mathrm{tr}} \to \infty} D(t_{\mathrm{tr}}) = 0,$$

the aforementioned assumption could be experimentally substantiated, although, so far, we did not use a very large training corpus (of several hours of speech) to show the validity of Eq. 3 also for very large t_{tr}. Furthermore, there seems to be a strong dependence of the parameters a and b on the weight α of Eq. 1, a behavior that is to be investigated in a future study.

At any rate, these considerations suggest to carefully select amount and nature of 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.

5. References

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

A. Black, R. Brown, R. Frederking, K. Lenzo, J. Moody, A. Rudnicky, R. Singh, and E. Steinbrecher. 2002. Rapid Development of Speech-to-Speech Translation Systems. In *Proc. of the ICSLP'02*, Denver, USA.

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. 2005a. TC-STAR: TTS Progress Report. Technical report.

A. Bonafonte, H. Höge, H. Tropf, A. Moreno, H. v. d. Heuvel, D. Sündermann, U. Ziegenhain, J. Pérez, and I. Kiss. 2005b. TC-Star: Specifications of Language Resources for Speech Synthesis. Technical report.

D. Chazan, R. Hoory, A. Sagi, S. Shechtman, A. Sorin, Z. Shuang, and R. Bakis. 2006. High Quality Sinusoidal Modeling of Wideband Speech for the Purposes of Speech Synthesis and Modification. In *Proc. of the ICASSP'06*, Toulouse, France.

H. Franco, J. Zheng, K. Precoda, F. Cesari, V. Abrash, D. Vergyri, A. Venkataraman, H. Bratt, C. Richey, and A. Sarich. 2003. Development of Phrase Translation Systems for Handheld Computers: From Concept to Field. In *Proc. of the Eurospeech'03*, Geneva, Switzerland.

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

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

1996. Methods for Subjective Determination of Transmission Quality. Technical Report ITU-T Recommendation P.800, ITU, Geneva, Switzerland.

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

M. Mashimo, T. Toda, H. Kawanami, H. Kashioka, K. Shikano, and N. Campbell. 2002. Evaluation of Cross-Language Voice Conversion Using Bilingual and Non-Bilingual Databases. In *Proc. of the ICSLP'02*, Denver, USA.

M. Pitz and H. Ney. 2005. Vocal Tract Normalization Equals Linear Transformation in Cepstral Space. *IEEE Trans. on Speech and Audio Processing*, 13(5).

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

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

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

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

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

D. Sündermann, H. Höge, A. Bonafonte, H. Ney, and J. Hirschberg. 2006b. Text-Independent Cross-Language Voice Conversion. In *Submitted to the Interspeech'06*, Pittsburgh, USA.

D. Sündermann, H. Höge, and T. Fingscheidt. 2006c. Breaking a Paradox: Applying VTLN to Residuals. In *Proc. of the ITG'06*, Kiel, Germany.

A. Waibel, A. Badran, A. Black, R. Frederking, D. Gates,

A. Lavie, L. Levin, K. Lenzo, L. Mayfield, J. Reichert, T. Schultz, D. Wallace, M. Woszczyna, and J. Zhang. 2003. Speechalator: Two-Way Speech-to-Speech Translation in Your Hand. In *Proc. of the NAACL'03*, Edmonton, Canada.

H. Ye and S. Young. 2004. High Quality Voice Morphing. In *Proc. of the ICASSP'04*, Montreal, Canada.