TC-STAR: Evaluation Plan for Voice Conversion Technology*

David Sündermann[1,2], Antonio Bonafonte[1], Helenca Duxans[1], Harald Höge[2]
[1] Universitat Politècnica de Catalunya, 08034 Barcelona, Spain,
e-mail: {suendermann, antonio, hduxans}@gps.tsc.upc.edu
[2] Siemens AG, 81739 Munich, Germany, e-mail: harald.hoege@siemens.com

Introduction

Voice conversion is the adaptation of the characteristics of a source speaker's voice to those of a target speaker [1]. When evaluating voice conversion technology, generally, we have two questions in mind:

- Does the technique change the speaker identity in the intended way?
- How is the overall sound quality of the converted speech?

The answers can be found applying subjective and objective error criteria. The former is based on listening tests. The latter expresses the distance between the converted speech and corresponding reference speech of the target speaker. However, our experience shows that the objective evaluation of voice conversion technology features severe shortcomings [2, 3]. Consequently, in this paper, we develop a plan limited to subjective measures that is to be applied within the scope of the European speech-to-speech translation project TC-STAR [4].

The Corpora

In TC-STAR, both conventional intralingual and cross-language voice conversion [5] are to be investigated. The considered languages are English, Spanish and Mandarin, the combinations for cross-language voice conversion are English-Spanish and English-Mandarin.

The Training Corpus

To generate the training corpus, for each language combination, four bilingual speakers (two female and two male) produce about one hour of speech of both covered languages. The read contents are based on parallel texts taken from parliamentary speeches, cf. [6].

The Evaluation Corpora

For subjective evaluation, we found that none of the conventional procedures provides the information required for completely answering the first above question [3]. Therefore, we suggest an evaluation method to be used in TC-STAR that, in some respects, is based on a proposal of Kain and Macon [7]. Having a look at state-of-the-art voice conversion technology, we note that most of the systems only transform vocal tract and excitation [8], whereas some approaches aim at transforming the speaker-dependent prosody as well [9]. To be applicable to both kinds of systems, we propose to create two separate evaluation corpora that exclude or include prosody conversion, respectively.

The Evaluation Corpus Excluding Prosody. In order to achieve a similar prosody of all involved speakers, we apply an extension of the 'mimic' approach presented in [7]: At first, we ask one template speaker of each of the considered languages to produce 200 sentences (about 20 min of speech). Then, for the Spanish and Mandarin corpus, we invite eight corpus speakers (four female and four male), four of them are the bilingual speakers who generated the training corpus mentioned above. For the English corpus, we have twelve speakers (six male and six female), including the eight bilingual speakers from the training corpus. In order to build consistent corpora of all languages, for the evaluation of English intralingual conversion, we choose those four speakers that have the most native-like pronunciation. For cross-language conversion to English, we take those speakers that have the source language as mother tongue.

150 of the sentences of each speaker are provided as adaptation corpus, whereas the remaining 50 sentences are the evaluation corpus.

During the recording session, the corpus speakers listen to a sentence of the template speaker and, then, try to mimic its timing and accentuation pattern and its pitch contour. Since, in general, the average fundamental frequency of the corpus speakers can differ essentially from that of the template speaker, the template utterances have to be adapted so that the corpus speakers feel comfortable mimicking the given pitch contour. Therefore, the template utterances are manipulated by means of a PSOLA technique that changes the fundamental frequency by adding a positive or negative speaker-dependent offset while keeping the speaking rate and the voice characteristics [10]. In particular, this adaptation is necessary when template and corpus speaker have different genders.

The Evaluation Corpus Including Prosody. Here, we expect the corpus speakers to use their individual prosody, i.e., no template speaker is required.

Subjective Evaluation

The evaluation is carried out using a web interface. This makes possible that the subjects can perform the test from their home computer that has to be equipped with a high-speed internet connection, a standard sound card and closed headphones. For each language, between 15 and 20 subjects participate in the evaluation. In order to prevent the subjects from interpreting their decisions, they should not be familiar to the background of the test. In particular, they must not know the contents of this evaluation plan. I.e., ideal evaluation subjects are persons that do not have specific knowledge about speech processing at all.

The evaluation web page contains a clear instruction of what the subjects are to do, e.g.:

"We are analyzing differences of voices. For this reason, you are asked to identify if two samples come from the same person or not. Please, do not pay attention to the recording conditions or quality of each sample, only to the identity of the person.

*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

So, for each pair of voices, do you think they are

(1) definitely different,

(2) probably different,

(3) not sure,

(4) probably identical,

(5) definitely identical?"

Voice Identity Conversion

To keep the evaluation task as convenient and clear as possible, two speech samples are presented at a time. Each speech sample consists of 10 sentences that are randomly chosen from the evaluation corpus consisting of 50 sentences, cf. above. The subjects are not forced to listen to the complete sample but can stop the playback whenever they want. The samples of two compared voices are based on identical sentences, whereas, for each comparison, the randomization is executed anew to prevent the subjects from becoming bored. Each subject evaluates the same test, i.e., the randomizations are executed beforehand. The evaluated voice conversion system has to convert the determined 10 sentences from each of the four training corpus voices (source voices) to each of the four evaluation corpus voices, i.e., we have 16 voice pairs. During the evaluation, the subjects listen to 16 voice pairs consisting of the conversion results and the respective reference (target) speech. Besides, they have to rate the similarity of the unconverted voices, i.e., we have 16 more pairs that consist of the source speech and the reference. These 32 voice pairs are randomized, thus the subject does not know if he compares the converted voice with the source or the target.

Preparing the Test. As explained above, during the recording of the evaluation corpus excluding prosody, we adjust the pitch of the template speaker by adding a pitch offset in the way that the respective corpus speaker feels comfortable. To make the prosody as speaker-independent as possible, in the test, this offset is to be deducted. This is done by providing the evaluated voice conversion system with the values of the pitch offset of the source and target speaker for each considered pair of utterances in the test. As each voice conversion system should include a pitch modification facility, this pitch offset is to be taken into account when synthesizing the converted speech. When comparing unconverted source and target utterances, the mean pitch of the source speech is adapted to that of the target speech by means of a PSOLA technique. A deterioration of the speech quality could be accepted as the subjects are asked to ignore it when evaluating the voice identity conversion.

Voice Conversion Score. In order to compare the performance of different voice conversion systems or to control a system's progress from one evaluation to the next, we define a voice conversion score that has to have similar properties as the mean opinion score used for quality assessment, cf. below. Since the performance of the conversion highly depends on the difference of the involved voices (source and target), this score should take into account both the distance between the converted and the target voice and that between source and target voice. When applying objective criteria to evaluate the voice conversion performance, one uses the ratio between both distances [2], however, here, the distances are normalized to between 0.0 and 1.0. A respective ratio for our subjective evaluation that keeps the score definition introduced above looks as follows:

$$s = 5 - \frac{20 - 4s(converted, target)}{5 - s(source, target)} .$$

Note that

- this equation becomes 1.0 if $s(converted, target) = s(source, target)$, i.e., if the conversion showed no progress.
- If $s(converted, target) < s(source, target)$, one should set $s = 1.0$ per definition.
- If $s(converted, target) = s(source, target) = 5$, the sample should not be counted.

The final voice conversion score is the mean over all considered samples of all involved subjects.

Overall Speech Quality

Since it is widely used in telecommunications, for measuring the quality of the converted speech, we apply a mean opinion score test [11]. The listeners are asked to assess certain sentences according to the following scale: (1) bad; (2) poor; (3) fair; (4) good; (5) excellent. The mean opinion score is the arithmetic mean of all subjects' individual scores.

Test Definition. To determine the best achievable conversion quality, the eight voices contained in the training and in the evaluation corpus are also considered. For the test, they are mixed up with the 16 conversion outputs.

References

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

[2] 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.

[3] D. Sündermann, "Voice Conversion: State of the Art and Future Work," in *Proc. of the DAGA'05*, Munich, Germany, 2005.

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

[5] 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.

[6] H. Höge, A. Bonafonte, H. v. d. Heuvel, A. Moreno, H. S. Tropf, D. Sündermann, and U. Ziegenhain, "TC-STAR Deliverable D8a: Specifications of Language Resources for Speech Synthesis, Tech. Rep. Draft – v2.23, 2005.

[7] 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.

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

[9] D. Rentzos, S. Vaseghi, Q. Yan, and C.-H. Ho, "Voice Conversion through Transformation of Spectral and Intonation Features," in *Proc. of the ICASSP'04*, Montreal, Canada, 2004.

[10] W. B. Kleijn and K. K. Paliwal, *Speech Coding and Synthesis*. Amsterdam, Netherlands: Elsevier Science B.V., 1995.

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