DONOR SELECTION FOR VOICE CONVERSION

Oytun Turk and Levent M. Arslan

Electrical and Electronics Eng. Dept., Bogazici University, Bebek, 34342, Istanbul, Turkey
R&D Dept., Sestek Inc., ARI-1 Teknopark Binasi, 34469, Istanbul, Turkey
phone: + (90 212) 286 25 44, fax: + (90 212) 286 25 47, email: <oytun, levent>@sestek.com.tr
web: www.busim.boun.edu.tr   www.sestek.com.tr

ABSTRACT

Voice conversion techniques enable the transformation of a source speaker's voice to that of a target speaker's automatically. The performance of any voice conversion algorithm depends on the source-target pair chosen. This study focuses on the problem of source speaker (donor) selection from a set of available speakers that will result in the best quality output for a specific target speaker's voice. A voice conversion database of 20 speakers (10 male, 10 female) is collected. 180 conversions that cover all male-to-male and female-to-female voice conversion combinations are performed using a codebook mapping based method. A listening test is performed in order to determine the subjective scores for similarity of the output to the target speaker's voice and the output quality. The results show that selecting the appropriate donor improves voice conversion performance significantly. Preliminary analysis is performed for automatic donor selection with multilayer perceptrons.

1. INTRODUCTION

Voice conversion is aimed at the automatic transformation of a source speaker's voice to a target speaker's voice. Although several algorithms are proposed for this purpose [1], [2], [3], [4], none of them can guarantee equivalent performance for different source-target speaker pairs. The dependence of voice conversion performance on the source-target speaker pairs is a disadvantage for practical applications. However, in most of the cases, the target speaker is fixed, i.e. the voice conversion application aims to generate the voice of a specific target speaker and the source speaker can be selected from a set of candidates. As an example, consider a dubbing application that involves the transformation of an ordinary voice to a celebrity's voice. In this case, choosing an appropriate source speaker (donor) among a set of candidates can enhance the output quality significantly. However, it is time-consuming and expensive to collect the entire training database from all candidates, perform conversions, and obtain the subjective decisions of the customer on the output quality. Another solution for donor selection might be to employ objective criteria in the selection process by comparing acoustical features obtained from a limited number of source and target utterances without actually performing conversions. In this case, the main issue becomes finding a relationship between the objective criteria and the output quality.

Considering the difficulties in selecting an appropriate source speaker for a specific target speaker, this study focuses on:

- the design and practice of a subjective listening test for the evaluation of voice conversion outputs among a large number of source-target speaker pairs
- the preliminary analysis of source-target speaker acoustical characteristics for automatic donor selection using multilayer perceptrons (MLPs)

Voice conversion has been a popular topic in speech processing research [1], [2], [3], [4]. In this study, STASC is employed which is a codebook mapping based algorithm proposed by one of the authors in [2]. STASC employs adaptive smoothing of the transformation filter to reduce discontinuities and results in natural sounding and high quality output. It is being used in commercial applications for international dubbing, singing voice conversion, and creating new TTS voices.

Figure 1 shows an overview of the proposed method for donor selection. Section 2 starts with the description of the voice conversion database collected and the subjective listening test designed for the evaluation of different source speakers to generate a target voice using voice conversion. Subjective listening test results are discussed next. Section 3 describes the acoustical features employed for estimating the objective distances between the source and target acoustical spaces. Preliminary analysis is performed for automatic donor selection using MLPs. In Section 4, the proposed automatic donor selection algorithm is evaluated. Finally, the study is concluded with a discussion in Section 5.
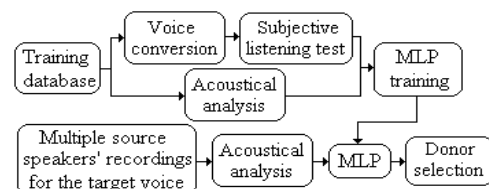

Figure 1. Overview of the donor selection method.

2. SUBJECTIVE LISTENING TEST

A subjective listening test is performed in order to obtain the subjective scores for voice conversion outputs of 180 source-target pairs in terms of similarity to the target voice and MOS quality. The results of this test are used in the next

section in preliminary analysis for automatic donor selection.

2.1. Database

The voice conversion database consisted of 20 utterances (18 training, 2 testing) from 10 male and 10 female native Turkish speakers recorded in an acoustically isolated room. The utterances were natural sentences describing the room like "There is a grey carpet on the floor". The electroglottograph (EGG) recordings were collected simultaneously. One of the male speakers was selected as the reference speaker and the remaining speakers were told to mimic the timing of the reference speaker as closely as possible. This helps to improve automatic alignment performance in voice conversion significantly.

2.2. Codebook Mapping Based Voice Conversion

In this study, STASC is employed for voice conversion. It is a two-stage codebook mapping based algorithm. In the training stage, the mapping between the source and target acoustical parameters is modelled. In the transformation stage, a novel method is employed to match the source speaker acoustical parameters with the source speaker codebook entries on a frame-by-frame basis and the target acoustical parameters are estimated as a weighted average of the target codebook entries. The weighting algorithm reduces discontinuities significantly. Details of STASC can be found in [2].

2.3. Method

We have considered male-to-male and female-to-female conversions separately in order to avoid quality reduction due to large amounts of pitch scaling required for inter-gender conversions. Each speaker was considered as the target and conversions were performed from the remaining nine speakers of the same gender to that target speaker. Therefore, the total number of source-target pairs was 180 (90 male-to-male, 90 female-to-female).

Twelve subjects were presented with the source, target, and transformed recording and were asked to provide two subjective scores for each transformation: similarity of the transformation output to the target speaker's voice (S-score) and the MOS quality of the voice conversion output (Q-score). S-score was in the range 1-10, 1 corresponding to the case when the transformation output does not sound like the target speaker at all, and 10 corresponding to the case when the output sounds exactly like the target speaker. The Q-score corresponded to the standard MOS scale for sound quality: 1=Bad, 2=Poor, 3=Fair, 4=Good, 5=Excellent.

2.4. Results

Figures 2 and 3 show the average S-scores for all source-target speaker pairs. For male pairs, highest S-scores are obtained when the reference speaker was the source speaker. Therefore, the performance of voice conversion is enhanced when the source timing matches the target timing better in the training set. Excluding the reference speaker, the source speaker that results in the best voice conversion performance varies as the target speaker varies. Therefore, our fundamental hypothesis that the performance of the voice conversion algorithm is dependent on the specific source-target pair chosen is supported. The last rows of Figures 2 and 3 show that some source speakers are not appropriate for voice conversion as compared to others, i.e. male source speaker no. 4

and female source speaker no. 4. The last columns in Figure 2 and 3 indicate that it is harder to generate the voice of specific target speakers, i.e. male target speaker no. 6 and female target speaker no. 1. Figures 4 and 5 show the average Q-scores.
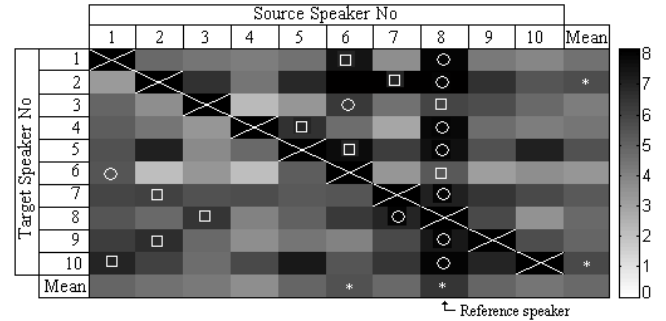

Figure 2. Average S-scores for all male source-target pairs. The two highest average scores for each source-target pair are marked with a small circle (highest) and a rectangle (second highest). An asterix is used for indicating the source and target speakers that have the highest average scores.
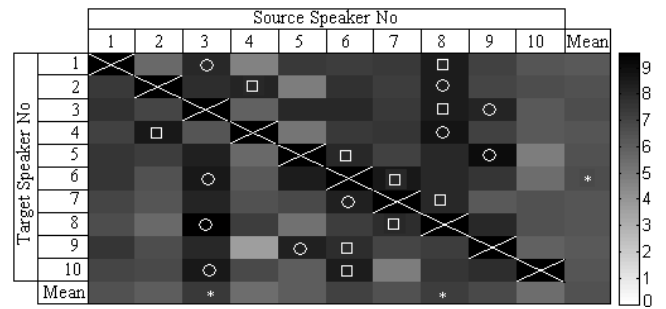

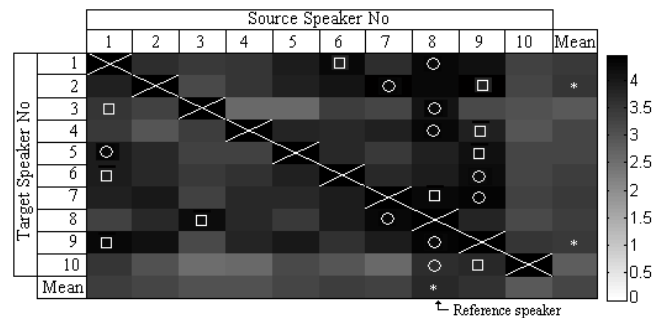**Figure 3.** Average S-scores for all female source-target pairs.

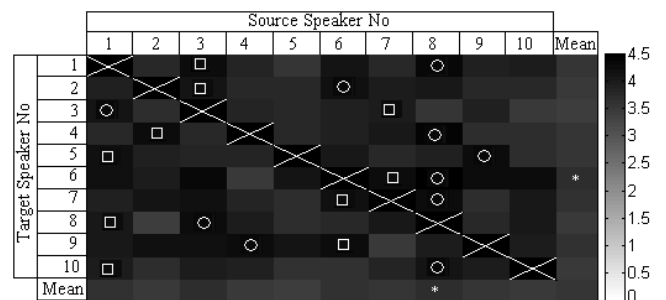**Figure 4.** Average Q-scores for all male source-target pairs.

Figure 5. Average Q-scores for all female source-target pairs.

3. AUTOMATIC DONOR SELECTION

We have used a set of acoustical features that are relevant for describing the differences among speakers. The aim was to determine the objective distances of different acoustical features for a source-target speaker pair and use these distances in estimating the subjective quality of the voice conversion output. Prior to acoustical analysis all recordings were phonetically labelled using HTK [5], the EGG signals of sustained vowel /aa/ are analysed and pitch marks are determined, pitch and energy contours are extracted, and corresponding frames are determined between each source and target utterance from the phonetic labels.

The following acoustical features and distances were used for comparing source-target speaker acoustical characteristics:

- Vocal Tract: LSFs are computed on a frame-by-frame basis using a linear prediction order of 20 at 16 KHz. The distance, d, between two LSF vectors is computed using:

$$d = \sum_{k=1}^{P} h_k |w_{1k} - w_{2k}| \qquad (1)$$

$$h_k = \frac{1}{argmin(|w_k - w_{k-1}|, |w_k - w_{k+1}|)} \; for \; k = 1, ..., P \quad (2)$$

where w_{1k} is the kth entry of the first LSF vector, w_{2k} is the kth entry of the second LSF vector, P is the prediction order, and h_k is the weight of the kth entry corresponding to the first LSF vector [2].

- Pitch: f0 values are computed using a standard auto-correlation based pitch detection algorithm.
- Duration: Phoneme, word, utterance, and inter-word silence durations are calculated from the phonetic labels.
- Energy: Frame-by-frame energy is computed.
- Spectral Tilt: The slope of the least-squares line fit to the LP spectrum (prediction order 2) between the dB amplitude value of the global spectral peak and the dB amplitude value at 4 KHz is used.
- Open Quotient (OQ): For each period of the EGG signals, OQ is estimated as the ratio of the positive segment of the signal to the length of the signal as shown in Figure 6.
- Jitter: Average period-to-period variation of the fundamental pitch period, T_0, excluding unvoiced segments in the sustained vowel /aa/ is computed using:

$$J = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |T_0(i) - T_0(i+1)|}{\frac{1}{N}\sum_{i=1}^{N} T_0(i)} \qquad (3)$$

- Shimmer: Average period-to-period variation of the peak-to-peak amplitude, A, excluding unvoiced segments in the sustained vowel /aa/ is computed using:

$$S = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |A(i) - A(i+1)|}{\frac{1}{N}\sum_{i=1}^{N} A(i)} \qquad (4)$$

- Soft Phonation Index (SPI): Average ratio of the lower-frequency harmonic energy in the range 70-1600 Hz to the harmonic energy in the range 1600-4500 Hz is computed.
- H1-H2: Frame-by-frame amplitude difference of the first and second harmonic in the spectrum is estimated from the power spectrum [6].
- EGG Shape: A simple, three parameter model to characterize one period of the EGG signals is used as shown in Figure 6 where α is the slope of the least-squares (LS) line fitted from the glottal closure instant to the peak of the EGG signal, β is the slope of the LS line fitted to the segment of the EGG signal when the vocal folds are open, and γ is the slope of the LS line fitted to the segment when the vocal folds are closing.
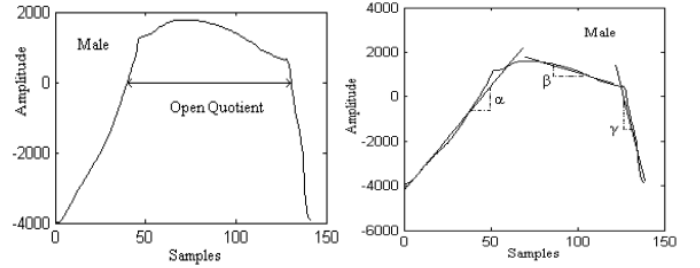

Figure 6. OQ estimation from EGG (left), simple model for EGG shape for a male speaker (right).
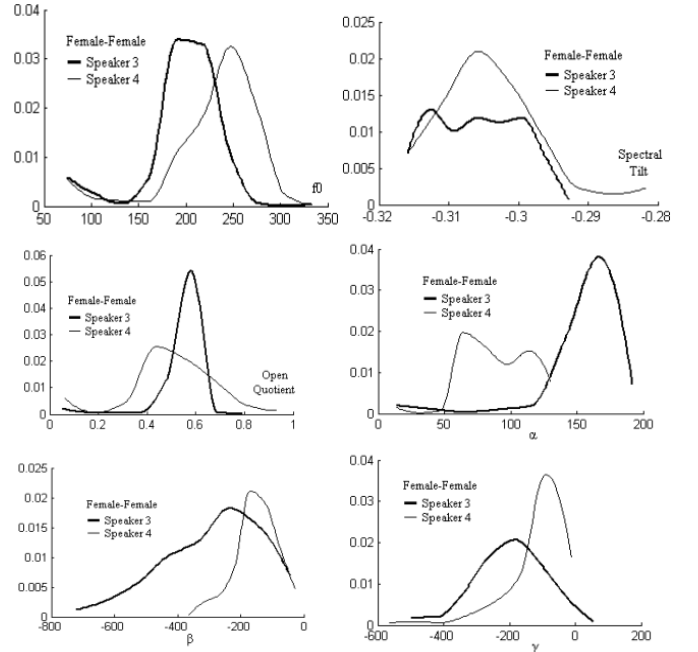

Figure 7. Histograms of different acoustical features.

Figure 7 shows histograms of different acoustical features. We have used the Wilcoxon rank-sum test to compare distributions of acoustical features for a given source-target speaker pair except the LSFs for which we already have the distance measure described above. The rank-sum test is a nonparametric alternative to the two-sample t-test, which is

valid for data from any distribution and is much less sensitive to the outliers as compared to the two-sample t-test [7]. It reacts not only to the differences in the means of distributions but also to the differences between the shapes of the distributions. The lower is the rank-sum value, the closer are the two distributions under comparison.

The rank-sum values and the statistics of LSF distances (mean and standard deviation) for each source-target pair are used as the input to an MLP with a single hidden layer. The S-scores and the Q-scores were set as the output of the MLP in order to estimate the subjective scores from a set of objective measures. For each source-target speaker pair, six utterances and two recordings of the sustained vowel /aa/ are used for calculating the objective measures. Figure 8 shows the flowchart of the automatic donor selection algorithm.
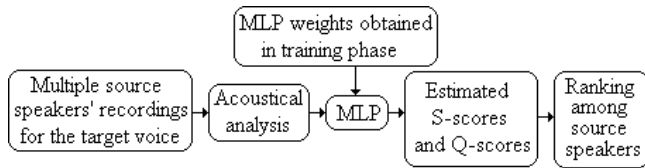

Figure 8. Flowchart of the donor selection algorithm.

4. EVALUATIONS

The performance of the proposed algorithm for donor selection is evaluated using 10-fold cross validation. For this purpose, two male and two female speakers are reserved as the test set. Two male and two female speakers are reserved as the validation set. The objective distances among the remaining male-male pairs and female-female pairs are used as the input to the MLP and the corresponding subjective scores as the output. After training, the subjective scores are estimated for the target speakers in the validation set and the error for the S-score and the Q-score is calculated. The error on each cross-validation step is defined as the absolute difference between the MLP's decision and the subjective test results:

$$E_S = \frac{1}{T} \sum_{i=1}^{T} |S_{SUB}(i) - S_{MLP}(i)| \qquad (5)$$

$$E_Q = \frac{1}{T} \sum_{i=1}^{T} |Q_{SUB}(i) - Q_{MLP}(i)| \qquad (6)$$

where T is the total number of source-target pairs in the test, $S_{SUB}(i)$ is the subjective S-score for the ith pair, $S_{MLP}(i)$ is the S-score estimated by the MLP for the ith pair, $Q_{SUB}(i)$ is the Q-score for the ith pair, and $Q_{MLP}(i)$ is the Q-score estimated by the MLP for the ith pair. E_S denotes the error in the S-scores and E_Q denotes the error in the Q-scores. The two steps described above are repeated 10 times by using different speakers in the validation set. The average cross-validation errors are computed as the average of the errors in the individual steps. Finally, the MLP is trained using all the speakers except the ones in the test set and the performance is evaluated on the test set. The results are shown in Table 1.

We are currently performing tests on decision trees trained with the ID3 algorithm to investigate the relationship between the subjective test results of Section 2 and the acousti-

cal distance measures of Section 3. As a preliminary result, a decision tree trained with data from all source-target speaker pairs distinguishes male source speaker no. 3 from the others by using only H1-H2 characteristics. The low subjective scores obtained when he is used as a target speaker indicate that it is harder to generate this speaker's voice using voice conversion. This speaker had significantly lower H1-H2 and f0 as compared to the rest of the speakers as correctly identified by the decision tree. Therefore, decision trees might provide further information that is not available in the case of MLPs for modifying the voice conversion algorithm to produce significant characteristics of the target voice in a better fashion.

Scores	10-fold Cross Validation		Test	
	E_S	E_Q	E_S	E_Q
Mean	0.83	0.21	0.77	0.18

Table 1. Results for 10-fold cross-validation and testing the MLP based automatic donor selection algorithm.

5. CONCLUSIONS

In this study, an automatic donor selection algorithm is proposed which estimates the subjective voice conversion output quality from a set of objective distance measures between the source and target speaker's acoustical features. The algorithm learns the relationship of the subjective scores and the objective distance measures through nonlinear regression with an MLP. Once the MLP is trained, the algorithm can be used in the selection or ranking of a set of source speakers in terms of the expected output quality for transformations to a specific target voice.

REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. of the IEEE ICASSP 1988*, pp. 565-568.

[2] L. M. Arslan, "Speaker transformation algorithm using segmental codebooks," *Speech Communication 28*, pp. 211-226, 1999.

[3] E. Moulines and Y. Sagisaka, (Eds.), "Voice conversion: state of the art and perspectives," *Special Issue of Speech Communication*, vol 16 (2), 1995.

[4] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Speech and Audio Proc.*, vol. 6, no. 2, pp. 131-142, 1998.

[5] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, "Large vocabulary continuous speech recognition using HTK," in *Proc. of the IEEE ICASSP 1994*.

[6] H. M. Hanson, "Individual variations in glottal characteristics of female speakers," in *Proc. of the IEEE ICASSP 1995*, pp. 772-775.

[7] C. J. Wild and G. A. F. Seber, Chance Encounters: A First Course in Data Analysis and Interference. John Wiley & Sons, Inc., 1999.