

# Subjective Evaluations for Perception of Speaker Identity Through Acoustic Feature Transplantations

Oytun Turk

Levent M. Arslan

Electrical and Electronics Eng. Dept.,  
Bogazici University, Istanbul, Turkey  
<turkoytu,arslanle>@boun.edu.tr

## Abstract

Perception of speaker identity is an important characteristic of the human auditory system. This paper describes a subjective test for the investigation of the relevance of four acoustic features in this process: vocal tract, pitch, duration, and energy. PSOLA based methods provide the framework for the transplantations of these acoustic features between two speakers. The test database consists of different combinations of transplantation outputs obtained from a database of 8 speakers. Subjective decisions on speaker similarity indicate that the vocal tract is the most relevant feature for single feature transplantations. Pitch and duration possess similar significance whereas the energy is the least important acoustic feature. Vocal tract + pitch + duration transplantation results in the highest similarity to the target speaker. Vocal tract + pitch, vocal tract + duration + energy and vocal tract + duration transplantations also yield convincing results in transformation of the perceived speaker identity.

## 1. Introduction

Perceiving the identity of people from their voices is an important characteristic of the human auditory system. We are able to recognize familiar voices by just listening to a few words or even a few phonemes. Many researchers have investigated the abilities and properties of the human auditory system in perception of speaker identity. In [1], the authors focus on objective measures and compare the performance of several objective descriptors for perception of speaker identity. They have shown that median pitch is correlated with speaker dissimilarity for both male and female speakers. They have also found correlations between vocal tract, glottal, and prosodic features and speaker dissimilarities.

In speech coding and speech synthesis applications, preserving speaker identity is an important performance criterion. Subjective testing methods are widely used for the evaluation of speaker recognizability in these systems. As an example, a subjective testing procedure is designed to evaluate the performance of a text-to-speech system in terms of speaker recognizability in [2]. Subjective listening tests are also used in the performance evaluation of voice conversion systems in [3], [4], and [5].

In our previous work, we have investigated the relevance of sub-band based spectral content for perception of speaker identity and shown that 1.0-1.8 KHz range was the most important frequency range [6]. In this study, we focus on four acoustic features (vocal tract, pitch, duration, and energy) and investigate the relevance of these features in the perception

process. The acoustic features are transplanted between two speakers using PSOLA based methods. The aim of these transplantations is to replace one or more of the acoustic features of a specific speaker with the acoustic features from another speaker. The transplantation outputs are used in a subjective test to determine their relative importance in perception of speaker identity.

We start with a description of PSOLA based acoustic feature transplantations in Section 2. Four types of single feature transplantations (vocal tract, pitch, duration and energy) and multi-feature transplantations are described in sub-sections 2.1-2.5 respectively. In Section 3, we describe the database and the procedure for the subjective listening test. The test results are given in Section 4. In Section 5, we conclude with a discussion on the results and future work.

## 2. Acoustic Feature Transplantations

In this section, we describe several PSOLA based methods for inter-speaker transplantation of acoustic characteristics. For all cases, we have two speakers – Speaker1 and Speaker2. Four different acoustic features are investigated in the transplantations: vocal tract, pitch, duration, and energy. We start with collecting the recordings of the same utterances from the source and the target speakers. These utterances are then phonetically labeled. Finally, the acoustical feature(s) of Speaker1 is (are) modified to match the acoustic feature(s) of Speaker2 using the phonetic alignment information in a PSOLA based framework.

As there are four acoustic features of concern, the number of all possible combinations of transplantations between Speaker1 and Speaker2 is 16. Two of these combinations correspond to original utterances which are denoted as “A” and “a” in shorthand notation in Table 1. The remaining 14 combinations can be obtained by performing only the following 7 transplantations in Table 1: “B”, “C”, “D”, “E”, “F”, “G”, and “H”. We repeat these transplantations by reversing the order of speakers to obtain the rest of the combinations in Table 1: “b”, “c”, “d”, “e”, “f”, “g”, and “h”. The corresponding time instant in the utterance of Speaker2 is determined by using the information from the labels and the analysis time instant in Speaker1 employing Equation 1. Note that, we assume a linear time-warping scheme within each phoneme. In Equation 1,  $i$  is the index of the current label,  $t_1^i$  is the time instant in Speaker1,  $t_2^i$  is the corresponding time instant in Speaker2,  $t_1^s$  is the start time of the  $i^{\text{th}}$  label in Speaker1,  $t_1^e$  is the ending time of the  $i^{\text{th}}$  label in Speaker1,  $t_2^s$  is start time of the  $i^{\text{th}}$  label in Speaker2, and  $t_2^e$  is the ending time of the  $i^{\text{th}}$  label in Speaker2.

$$t_2^i = \frac{(t_1^i - t_1^s)(t_2^e - t_2^s)}{t_1^e - t_1^s} + t_2^s \quad (1)$$

<sup>1</sup> This study was supported by Bogazici University 03A201 Research Fund Project.

Transplantation Type (from Speaker2 to Speaker1)	Shorthand Notation	Vocal Tract	Pitch Contour	Phonemic Durations	Energy Contour
Original (Speaker1)	A	1	1	1	1
Original (Speaker2)	a	2	2	2	2
Vocal tract	B	2	1	1	1
Pitch/Duration/Energy	b	1	2	2	2
Pitch	C	1	2	1	1
Vocal tract/Duration/Energy	c	2	1	2	2
Duration	D	1	1	2	1
Vocal tract/Pitch /Energy	d	2	2	1	2
Energy	E	1	1	1	2
Vocal tract/Pitch /Duration	e	2	2	2	1
Vocal tract/Pitch	F	2	2	1	1
Duration/Energy	f	1	1	2	2
Pitch/Energy	G	1	2	1	2
Vocal tract/Duration	g	2	1	2	1
Vocal tract/Energy	H	2	1	1	2
Pitch/Duration	h	1	2	2	1

Table 1: All possible acoustic feature transplantations between Speaker1 and Speaker2. “1” denotes Speaker1, “2” denotes Speaker2.

### 2.1. Vocal tract transplantation

Vocal tract transplantation is performed pitch-synchronously. First, the corresponding time instant in Speaker2 is calculated using Equation 1. Next, we perform LPC analysis on the utterances of Speaker1 and Speaker2 at the corresponding time instants and estimate the vocal tract parameters. We also calculate the excitation spectrum for Speaker1. Multiplying the excitation spectrum of Speaker1 with the vocal tract spectrum of Speaker2 generates the modified spectrum. The time domain output is obtained using inverse FFT and the overlap-add method. A vocal tract transplantation output is shown in Fig. 1.

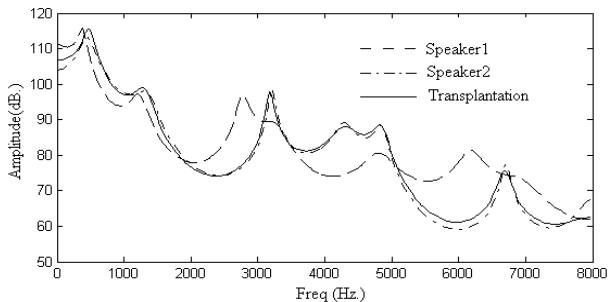


Figure 1: Vocal tract spectra for Speaker1, Speaker2, and transplantation output for a voiced phoneme.

### 2.2. Pitch Contour Transplantation

We determine the amount of pitch scaling required for transplanting the pitch contour of Speaker2 onto the utterance of Speaker1 by using the time alignment given by Equation 1. The instantaneous pitch-scaling ratio is given by the ratio of the instantaneous  $f_0$ -values of Speaker2 and Speaker1. We limit this ratio in the range  $[0.5, 2.0]$  in order to avoid exceptionally small or large pitch scaling factors. It is possible that voiced segments of the pitch contour of Speaker1 correspond to unvoiced segments in the pitch contour of Speaker2. In this case, unvoiced regions of the pitch contour

of Speaker2 are linearly interpolated using the neighboring voiced  $f_0$  values. The output of pitch contour transplantation is shown in Fig. 2.

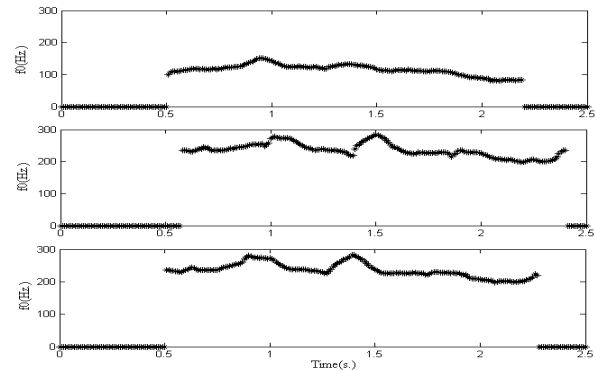


Figure 2: Speaker1 (Top), Speaker2 (Middle), and pitch contour transplantation output (Bottom) for a completely voiced utterance in Turkish.

### 2.3. Transplantation of Phonemic Durations

As the ratio of durations of source and target phonemes is variable across phonemes, one has to apply variable time-scaling ratios for the transplantation of phonemic durations. In most of the cases, we observe drastic changes in this ratio that makes duration modeling a very difficult problem to handle. In our case, as we have the exact correspondence between the durations of the source and target phonemes, all we need is a method for variable duration modification producing the target durations in an exact manner. Although PSOLA based algorithms generate high quality output for constant duration scaling ratios in low to medium ranges, the output quality degrades when the duration scaling ratio changes with time with sudden changes or a constant corresponding to high amounts of duration modification. We have employed an alternative method to overcome this problem. In our method, the vocal tract, pitch and energy characteristics of Speaker1 were transplanted onto the utterance of Speaker2 instead of transplanting the duration characteristics of Speaker2 onto Speaker1. Fig.3 shows an example of duration transplantation. We observe that the duration for the phoneme /a/ is 108 ms. for Speaker1 and 70 ms. for Speaker2. The duration of the source phoneme is reduced successfully to 70 ms as desired by duration transplantation.

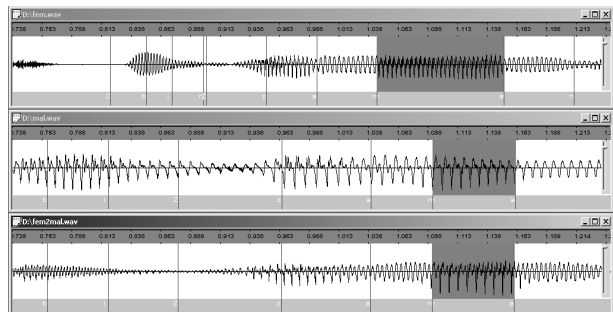


Figure 3: Phonetic labels and waveforms for Speaker1 (Top), Speaker2 (Middle), and duration transplantation from Speaker2 onto Speaker1 (Bottom).

By employing this method, it was possible to avoid variable duration scaling for phonemic duration transplantations. This method performs well because the modifications applied to the vocal tract, pitch and energy features do not distort the signal as much as applying duration scaling with a time-varying duration scaling ratio. However, by employing this type of duration transplantation, the source characteristics of Speaker2 are preserved except the pitch and the energy. So, the duration transplantation must in fact be considered as the complementary case of vocal tract + pitch + energy transplantation.

#### 2.4. Energy Contour Transplantation

The energy contour transplantation method is similar to the method used for pitch contour transplantation. Instantaneous values of the energy contours are used for obtaining the instantaneous energy scaling factor. The energy contours are smoothed before transplantation in order to reduce discontinuities at the phoneme boundaries.

#### 2.5. Multi-feature Transplantations

The rest of the transplantations are obtained by employing a combination of the four basic methods described above.

### 3. Subjective Test Design

We have used voice recordings of four male and four female speakers in our test database. The database contained 50 words and 29 short sentences in Turkish for each speaker. All gender combinations were employed for the transplantations: male-to-male, male-to-female, female-to-female, and female-to-male. 16 sentences and 16 words are selected randomly from the database for each speaker pair. In the listening tests, two sentences and two words are used as calibration utterances and reserved for assessing the reliability of the subjects' judgements. Each remaining sentence and word is used in one type of transplantation, so we obtain all possible transplantations using one sentence and one word. The subjects were provided with 128 utterance triples. Each triple contained two original utterances from two different speakers and one transplantation output. For reliability measurements, we have used original utterances of either the first or the second speaker as the third item as explained above. Ten subjects were used in the listening tests. Subjects were asked to make a speaker decision and assign a confidence score for each transplantation sample along with source and target speaker samples they heard. The speaker decision reflected subject's opinion on the identity of the speaker in the transplantation output. Subjects have simply decided whether the third item was uttered by "Speaker1", "Speaker2" or "None" of them. The subjects were also asked to provide a score reflecting how confident they are on their choice of speaker identity. The scoring scale ranged from 1 (least confident) to 5 (most confident). In the case that the transplantation output sounded similar to both speakers, the subjects were told to assign "None" as the speaker identity with a low confidence score. If the output sounds like a third speaker, listeners assigned a high confidence score.

We have observed that when the same recording is used over and over again, the ability of listeners to recognize the speaker identity degrades considerably. So, we have used different sentences and words for each type of transplantation

for a speaker pair in order to minimize the effect of the phonetic content of the signals and to emphasize on the capability of the listeners to recognize speakers. A graphical user interface was designed to carry out the tests. We have used voices of speakers that the subjects were familiar with because the aim was to evaluate the performance of the subjects when they had sufficient information on the identity of the speakers.

### 4. Results

The speaker identity decisions of the subjects are mapped onto a numerical scale to calculate the identity scores. The identity scores had three distinct values: 0.0, 0.5 and 1.0 corresponding to choices of "Speaker1", "None", and "Speaker2" respectively. As we reverse the order of Speaker1 and Speaker2 randomly, the decisions are preprocessed to reflect the similarity to Speaker2, i.e. if the first item in the triple belongs to Speaker2, the identity score is subtracted from 1.0. The confidence scores were normalized to unity. After these preprocessing steps, we estimate the mean and the inter-quartile ranges (IQRs) of the identity and confidence scores for the following cases: "Overall", "M→M", "M→F", "F→M", and "F→F". These cases correspond to different combinations in terms of speaker gender. "Overall" indicates that the statistics are calculated over all gender combinations. M denotes a male speaker, and F a female speaker. As an example, "M→F" denotes the case in which Speaker1 is a male and Speaker2 is a female. The mean scores for different cases are shown in Fig. 5 and Fig. 6. In each figure, we have two sub-plots for the identity and confidence scores respectively. In each sub-plot, we have different group of lines each corresponding to the cases described above regarding the genders. These groups are labeled on the x-axes by the corresponding case. Note that the case "Overall" is included in all sub-plots in order to compare the results of a specific case with the overall trends in which the gender of the speaker pairs are not considered. For each case, we have a group of 16 lines labeled at their top. These labels are the shorthand notations defined in the second column of Table 1. As an example consider the first group of lines (case "Overall") in the first sub-plot of Fig. 5. The third line (labeled as "B") corresponds to the case when the subject listens to a vocal tract transplantation output. The corresponding mean identity score is 0.32. On the second sub-plot of Fig. 5, we observe that the corresponding confidence score is 0.52. The first two lines (labeled as "A" and "a") serve as a basis for the evaluation of the reliability of the subjects. In both cases, the subjects are presented with an original recording of Speaker1 or Speaker2. Thus, the identity score should be 0.0 for "A" and 1.0 for "a". The confidence score should be 1.0 in both cases. We observe that the results match these values exactly indicating that the subjects have identified Speaker1 and Speaker2 perfectly for all pairs.

We have used IQRs as an indicator of the agreement between the responses of different subjects. Note that IQR is defined as the difference of the value which is greater than 75 percent of the data and the value which is greater than 25 percent of the data. So, it is an indicator of the spread of data like the standard deviation. If the IQR is close to 0.00, the data is not widespread. This is desired in our subjective test as it indicates that the subjects are in agreement for the corresponding transplantation. IQR values close to 1.00

indicate that the scores are widespread and the decisions of the subjects are not in agreement. We have considered the IQRs for the inferences that follow. We have also estimated the scores for words and sentences separately. However, we did not include figures like Fig. 5 and Fig. 6 for these cases. Instead, we have included remarks for the existing differences in the scores for words and sentences in the following paragraphs.

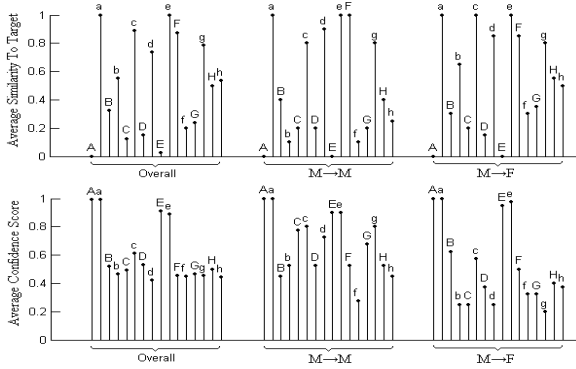


Figure 5: Subjective test results for all utterances.

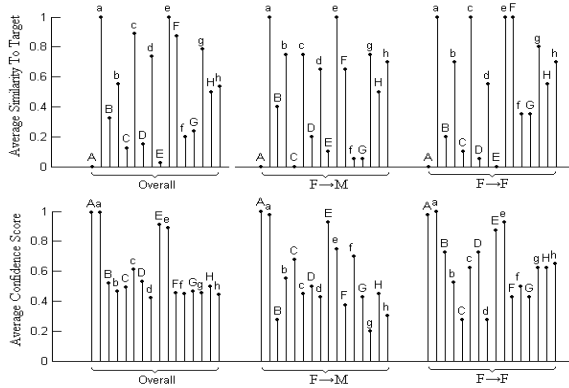


Figure 6: Subjective test results for all utterances.

In Fig. 5. and Fig. 6., we observe that the most convincing output in terms of similarity to Speaker2 is obtained by performing vocal tract + pitch + duration transplantation (e). The average confidence scores are also high for this type of transplantation. The IQRs are low (0.00 for both identity and confidence scores) so most of the subjects provided high scores for this type of transplantation. The complementary transplantation of the energy contour (E), was rated as the least convincing one regarding the similarity to Speaker2. Pitch contour (C) and pitch + energy contour (G) transplantations had lower scores with low IQRs. In the case of words, these scores were closer to the scores for energy contour transplantation (E) indicating that pitch information is not as important as in the case of sentences. This is expected because prosodic characteristics are more variable in sentences. Vocal tract + pitch contour (F), vocal tract + duration + energy (c) and vocal tract + duration (g) transplantations also had high scores. In the case of single feature transplantations, vocal tract is the most relevant feature. In most of the cases, pitch and duration had similar scores. The least relevant feature was the energy contour. Pitch transplantations had relatively higher scores when the

genders of the two speakers were different. The average scores also indicate that if any transplantation is assigned a higher score, the complementary transplantation gets a lower score. As an example, consider the energy contour transplantation (E) and the vocal tract + pitch + duration transplantation (e). We observe that the average similarity to the target speaker is lowest for energy contour transplantation (E) for the overall case in Fig. 5. The complementary case corresponding to vocal tract + pitch + duration transplantation was assigned a very high score close to 1.0. It is clear that if transplanting a subset of the features does not produce an output that sounds like the target speaker, transplanting the rest of the features or including more features will have a better chance.

## 5. Conclusions

In this study, we have designed a subjective listening test to evaluate the importance of four acoustic features in perception of speaker identity: vocal tract, pitch, duration, and energy. These acoustic features were transplanted using PSOLA based methods. The vocal tract + pitch + duration transplantation had the highest scores in terms of similarity to the target speaker. In the case of single feature transplantations, vocal tract was the most important feature. Pitch and duration had less importance than the vocal tract. Both features had similar importance, but the pitch characteristics were more important in the case that the genders of the speakers were different. The least important feature was the energy contour. We have also shown that complementary transplantations had complementary scores.

It would be beneficial to perform similar subjective tests in different languages. The methods described in this study can be used for the performance evaluation of automated conversion methods for acoustic features such as the methods employed in voice conversion. It is also possible to evaluate the importance of different acoustic features in perception of speaker identity using the methods described.

## 6. References

- [1] Necioğlu, B. F., Clements, M. A., Barnwell III, T. P., and Schmidt-Nielsen, A., “Perceptual Relevance of Objectively Measured Descriptors for Speaker Characterization”, in Proc. of the ICASSP 1998, Vol. 2, pp. 869-872, Seattle, WA, USA.
- [2] Sakamoto, M., and Saito, T., “Speaker Recognizability Evaluation of a VoiceFont-Based Text-to-Speech System”, in Proc. of the ICSLP 2002, pp. 2529-2532, Denver, CO, USA.
- [3] Arslan, L.M., “Speaker Transformation Algorithm Using Segmental Codebooks”, *Speech Communication* 28 (1999), pp. 211-226.
- [4] Kain, A. B., *High Resolution Voice Transformation*, Ph.D. Dissertation, OGI School of Science and Engineering at Oregon Health and Science University, 2001.
- [5] Turk, O., *New Methods For Voice Conversion*, M.S. Thesis, Bogazici University, 2003.
- [6] Ormanci, E., Nikbay, H., U., Turk, O., and Arslan, L. M., “Subjective Assessment of Frequency Bands for Perception of Speaker Identity”, in Proc. of the ICSLP 2002, pp.2581-2584, Denver, CO, USA.