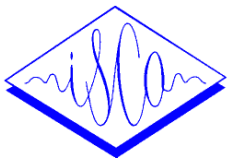


6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

SPECTRAL VOICE CONVERSION BASED ON UNSUPERVISED CLUSTERING OF ACOUSTIC SPACE

Masoud Geravanchizadeh

Institute of Communication Acoustics, Ruhr-University Bochum
D-44780 Bochum, Germany, e-mail: geravan@ika.ruhr-uni-bochum.de

ABSTRACT

Voice conversion systems aim at modifying a source speaker's speech so that it is perceived as if a target speaker had spoken it. Applying voice conversion techniques to a concatenative text-to-speech synthesizer allows for the personification of such systems, so that additional voices from a single source-speaker database can be produced quickly and automatically. This paper presents a new algorithm in which an effective and simple solution to the problem of voice conversion is suggested with the goal of maintaining high speech quality. Here, spectral conversion is performed by locally linear transformations, where the minimum mean square estimation (MMSE) method is used to compute the transformations. The acoustic features included in the conversion are vocal tract parameters, which are represented by log area ratio coefficients. Evaluation by listening tests shows that the proposed algorithm makes it possible to convert speaker individuality while maintaining high quality.

1. INTRODUCTION

Voice conversion is a technique to change speaker individuality; i.e., speech uttered by one speaker is modified to sound as if it had been articulated by another speaker. This topic has numerous applications, which include personification of text-to-speech systems, designing hearing aids appropriate to specific hearing problems, and so on.

The present work deals with the transformation of speech parameters between two different speakers at the segmental level. Specifically, we propose a new voice conversion algorithm that manipulates spectrum parameters related to the vocal tract. Our aim is to estimate a transformation function that maps the acoustic space of a source speaker to the acoustic space of the target speaker. As optimization criterion, minimum mean square estimation (MMSE) approach is used to compute the required transformations. Since the estimation is a linear method, the transformation functions can be represented by matrices.

In contrast to the codebook mapping methods, in which the speaker's spectrum space is represented by a fixed number of centroid vectors, which are finally used to define the mapping function, our proposed method considers the whole spectrum space in developing the conversion rules. This results in a continuous transformation of parameters between acoustical spaces of any pair of speakers. Accordingly, such unnatural spectral distortions in the transformation observed with codebook mapping approaches, as a vector jumps from one cluster to the other, are avoided and an improved quality of synthetic speech is achieved.

2. VOICE CONVERSION ALGORITHM

The proposed voice conversion algorithm is based on a set of linear transformation rules, which are selected according to the spectral features of a short-time signal in an operation phase. Our spectral transformation technique consists of the following steps: training phase and transformation-synthesis phase. The training phase is a process to generate conversion rules of spectrum parameters for each subspace obtained by vector quantization. In the transformation-synthesis phase, the conversion rules are applied to the the spectral features of the source speaker. Both steps are performed with a fixed frame rate.

2.1. Training procedure

As training corpus, we use a word set from the speech data of both speakers, which is analysed and the log area ratio (LAR) coefficients are extracted as spectrum parameters. The reason for choosing LARs is that these parameters are stable and have suitable spectral characteristics as well as good interpolation properties [1]. Here, we use a fixed frame rate to extract the required acoustical features. The next step consists of obtaining a mapping between acoustic spaces of source and target speakers. The approach taken in this work is based on using a standard unsupervised clustering technique in which the acoustic space of the source speaker is first divided into non-overlapping classes. Then, each class of feature vectors is associated with a transformation function, which can be obtained through some statistical procedures described below. The whole training procedure consists of the following steps:

- Partitioning the spectral space of source speaker
- Determining the corresponding partitions of target spectral space using DTW
- Calculating the transformation rules for each pair of corresponding partitions

Using the LBG algorithm [3], the feature space of the source speaker is first partitioned into non-overlapping classes. The next step in the training procedure involves obtaining the corresponding classes of the target spectral space. At this stage the same words from the speech corpus are used for classifying the feature space of the target speaker as used for making the source-speaker partitions. The idea is to use a quantized feature space of the source speaker in order to get a suitable classification of the target's feature space. Here, first both speakers utter the same word from the training material, which is analysed and the required features are extracted. Then source and target spectral vectors are aligned using a modified dynamic time warping algorithm [2]. The outcome of this DTW for each word is a succession of time-aligned vectors from the speech of both speakers. Now each time-aligned feature vector of the source speaker is mapped into the quantized feature space through vector quantization yielding a codebook index, idx_m. The index of the partition so obtained provides an address, where the corresponding vectors of the target speaker will be accumulated to form a partition. This procedure is repeated similarly for all other words in the training corpus.

The final step in the training process deals with generating the transformation rules specific to each class. Here again, a time-alignment of speech uttered by both source and target speaker is carried out using DTW algorithm. Before modelling the transformation rule for each class, however, it is necessary that the time-aligned feature vectors of source and target speaker are normalized with respect to the mean vector (centroid) specific to each class. This is accomplished by subtracting the corresponding centroid from each time-aligned spectral vector. The normalized vector pairs are then collected at a location in the conversion modul specified by the index idx_m. As before, the index idx_m is obtained by quantizing each time-aligned feature vector of source.

Assume $\tilde{\mathbf{X}}^{s,m} = \{\tilde{X}_j^{s,m}\}_{j=1}^{N_m}$ and $\tilde{\mathbf{Y}}^{s,m} = \{\tilde{Y}_j^{s,m}\}_{j=1}^{N_m}$ represent the set of all normalized feature vectors of the source and target speaker, respectively, which reside in the *mth* partition of the conversion modul. By using MMSE approach, the matrix $\mathcal{A}_m$, representing the linear regression transformation corresponding to this partition, is obtained as the solution to the minimization problem

$$\sum_{j=1}^{N_m} \| \tilde{Y}_j^{t,m} - \mathcal{A}_m \cdot \tilde{X}_j^{s,m} \|^2,$$

which is denoted by the following relation:

$$\mathcal{A}_m = \tilde{\mathbf{Y}}^{t,m} \cdot \tilde{\mathbf{X}}^{s,m^\dagger}, \quad (1)$$

where † denotes the pseudo-inverse operator.

2.2. Transformation procedure

The functional diagram of our proposed voice conversion system is illustrated in Figure 1. To cope with spectral transformation, we make use of the classical source-filter decomposition: The signal is split into a global spectral envelope which accounts for the resonant characteristics of the vocal tract transfer function together with the spectral characteristics of the glottal excitation and lip radiation ("spectral tilt") and a flattened source component, containing much of the prosodical information. Since the present work deals primarily with the transformation of spectral characteristics related to the vocal tract, it is necessary to pass the speech wave through a first-order pre-emphasis filter before the analysis to compensate for the spectral tilt in the speech signal.

According to the diagram, the tasks involved in the conversion procedure consist of time-scaling operation (Figure 1.a) and spectral transformation (Figure 1.b), which are discussed in more detail below.

Time-scale modification: To make the task of voice conversion simple and yet obtain a reasonable framework to evaluate the performance of spectral transformation, we simply copy the prosody of each target word during the transformation process. This is done in the time-scaling modul. Actually, time-scaling is a process to obtain an excitation signal through DTW for driving the synthesis modul which has the prosodic information of the target, but at the same time has the same number of frames as the source speaker. To accomplish this task, the warping path provided by DTW is considered as a piecewise linear function, in which one and only one frame from the target stream is assigned to each frame of the source signal.

Spectral transformation: The steps involved in the spectral transformation are as the following: The speech data of source speaker are first analysed at a fixed frame rate and LARs are extracted as transforming parameters. The next step consists in finding the partition to which each analysis vector belongs. This is done by finding the nearest codeword according to the squared-error distance measure. The result is a mean vector, $C_{A,m}$, and an index, idx_m, characterizing the location of the partition in the source acoustical space. The normalized vector is then obtained by substracting the mean vector $C_{A,m}$ from the analysis vector. The index idx_m determines an appropriate transformation related to this subspace, i.e. Matrix $\mathcal{A}_m$, which is multiplied with the normalized vector, yielding the normalized transformed vector, $\widetilde{X_j^{s \Rightarrow t}}$. At the same time, idx_m specifies

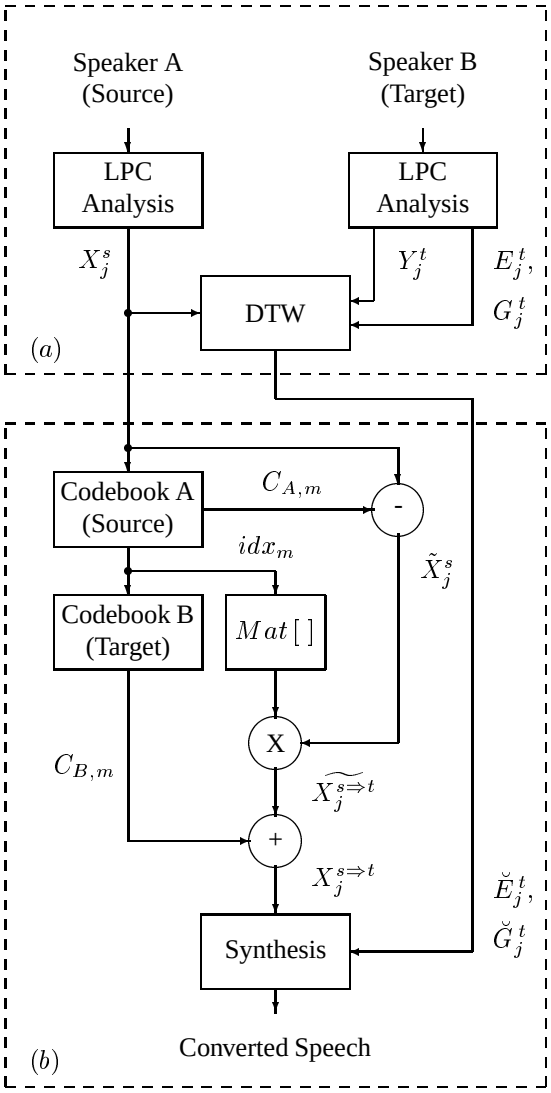


Figure 1: Spectral transformation from speaker A to speaker B

also a corresponding mean vector, $C_{B,m}$, from the target space, which is added to the new vector $\widetilde{X_j^{s\Rightarrow t}}$ to form the denormalized vector $X_j^{s\Rightarrow t}$. Finally, a new signal is synthesized from the modified excitation signal of the target and the transformed spectral features of the source. Since pre-emphasis has been used in the analysis step, de-emphasis must be applied at the output of the synthesis filter.

3. PERFORMANCE EVALUATION

We conducted formal listening tests and examined spectrograms to assess the performance of our proposed voice conversion algorithm. First, in the training phase, conversion rules for both speakers were generated using isolated word utterances. Then, in the transformation-synthesis phase, 62 uttered words from the training corpus were converted according to the steps discussed above.

Learning words	62
Sampling frequency	16 kHz
LPC analysis	autocorrelation method
LPC analysis order	20
Frame length	30 msec
Frame shift length	10 msec
Window function	Hanning window
Clustering distance measure	squared-error distance
Transforming parameters	log area ratio coefficients
Size of clusters	64

Table 1: The analysis conditions to generate the transformation rules

The speech database [4] of the BMBF project "SPINA" has been used as the training material. For each speaker, there are five utterances of each word, of which the first four repetitions are designated as training tokens, whereas the fifth is used for testing. The language is German. Table 1 shows the experimental conditions.

3.1. Preliminary evaluation experiments

In order to evaluate the proposed algorithm, the accuracy of formant tracking was first examined. Figure 2 illustrates one example of the speech spectrograms. Judging from the results, the proposed algorithm shows very good performance in terms of F1, F2 and F3 tracking. In the converted speech (Figure 2.b), it is clearly observed that F3 has moved downward from the source speech (Figure 2.a) to F3 of the target speech (Figure 2.c). Moreover, the formant-frequency change is quite smooth.

3.2. Evaluation by listening tests

To evaluate the overall performance of spectral conversion, four kinds of hearing tests were carried out based on a 62 word database of three male and two female speakers, five versions each. The first and second experiments concern with the conversion between male and female speakers, whereas the other two experiments deal with the male-to-male and female-to-female conversion respectively. The listening tests were designed to evaluate the accuracy of speaker-individuality conversion using the following procedure: Stimuli (1) and (2) were the speech converted by the proposed algorithm or a synthesized speech with the spectral parameters of source, but the excitation signal of target. Stimulus R was the target speaker's speech. Listeners were asked to select either (1) or (2) as being closest to R. Table 2 shows the preference rate of our proposed method.

The voice conversion performance depended upon the distance between the source speaker and the target speaker. In terms of male-to-male and female-to-female conversion, the spectrum distortion between them were small, which

Source	Target	Preference rate (%)
female $_{(51)}$	male $_{(01)}$	95.21
male $_{(01)}$	female $_{(51)}$	94.08
male $_{(29)}$	male $_{(28)}$	87.29
female $_{(75)}$	female $_{(51)}$	84.67

Table 2: Results of the hearing tests

means that their speech quality is very close in nature. This could be a reason why the voice-conversion performance was lower for these pairs.

4. CONCLUSION

A new voice-conversion algorithm has been proposed. The algorithm is characterized by its use of piecewise-linear conversion rules to precisely modify spectral parameters related to the vocal tract. Listening tests showed that the proposed algorithm makes it possible to convert speaker individuality while maintaining high quality. The algorithm makes it possible to convert static characteristics (spectrum envelopes). However, dynamic characteristics can not be converted. It is considered essential, however, that dynamic properties of speech be examined and included in the conversion process. Future work will therefore have to focus on methods for representing transitional or dynamic features to be used in the voice conversion system.

5. ACKNOWLEDGMENT

I thank Professor Jens Blauert, head of the Institute of Communication Acoustics, for the opportunity to perform this project under his supervision.

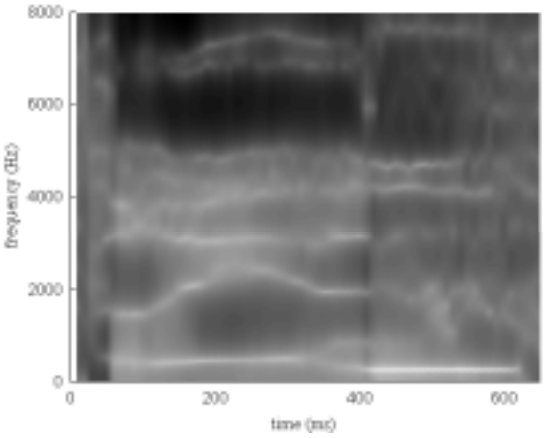
6. REFERENCES

[1] B.S. Atal. Efficient coding of LPC parameters by temporal decomposition. *Proc. Internat. Conf. Acoust. Speech Signal Proc.*, pages 81-84, 1983.

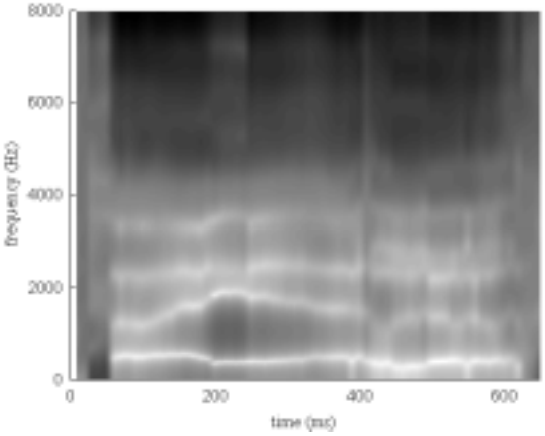
[2] M. Geravanchizadeh and M. Schaaf. Eine modifizierte nichtlineare Zeitachsentransformation für die spektrale Transformation von Stimmen. *Zehnte Konferenz Elektronische Sprachsignalverarbeitung, ESSV'99*, pages 72-77, 20-22 September 1999. Görlitz, Germany.

[3] Y. Linde, A. Buzo, and R.M. Gray. An Algorithm for vector Quantiser Design. *IEEE Trans. on Communication*, COM-28:84-95, January 1980.

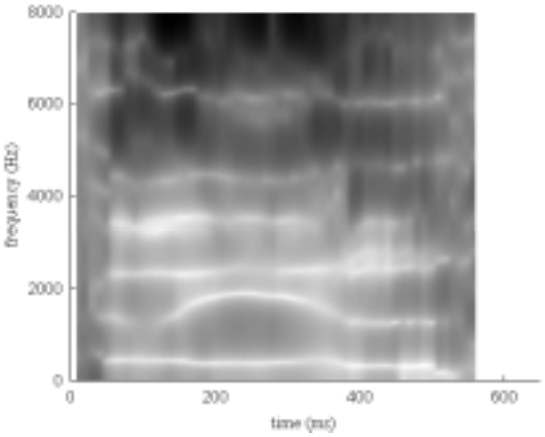
[4] SPINA. Sprachverstehen in neuronaler Architektur. Förderkennziffer 413-4001-01 IN 108, FRG, Verbundprojekt des Bundesministers für Bildung und Forschung (BMBF), 1991.



a) Source spectrum



b) Modified spectrum



c) Target spectrum

Figure 2: Comparison of speech spectrograms; female-to-male spectral transformation for the German word "drehen"