Voice Conversion for Unknown Speakers

Hui Ye and Steve Young

Cambridge University Engineering Department
Trumpington Street, Cambridge, England, CB2 1PZ
hy216@eng.cam.ac.uk, sjy@eng.cam.ac.uk

Abstract

Voice conversion is a technique for modifying a source speaker's speech to sound as if it was spoken by a target speaker. The conventional solutions to this problem are based on training and applying conversion functions which require a substantial amount of training data from both the source and the target speaker. In this paper, we present a voice conversion technique that requires no pre-existing training data from the source speaker. This new approach uses a speech recognizer to index the target training data so that each unknown source frame can be used to retrieve similar frames from the target database. The retrieved frames are then used to estimate conversion functions in a similar way to conventional methods. The paper presents both objective and subjective evaluations of the method. It also explores a number of variants including the contrast between using single and multiple transforms, and between the cases where the content of the source speech is known or unknown. The overall conclusion of the paper is that the method presented can result in identification of the target speaker with as little as a single sentence of source data to transform, however, knowledge of the source orthography is needed to attain a close similarity.

1. Introduction

The purpose of voice conversion is to transform a source speaker's speech to sound as if it was produced by a target speaker. In general, a voice conversion system includes two parts, the training procedure and the transformation procedure. In the training procedure, the source and the target speech training data are parameterised and then a conversion function is trained to capture the relationship between them. Given the trained conversion function, arbitrary speech from the source can then be converted to sound like the target. Ideally, the conversion function should provide near lossless speaker transformation without causing distortions and discontinuities in the converted speech. The ability to do this will depend on many factors but sufficient training data to robustly estimate the conversion functions is a key requirement. In previous research, a variety of approaches have been developed to implement the conversion function, such as Codebook Mapping [1], Linear Transformations [2, 3] and Continuous Probabilistic Transformations [4]. However, all these methods assume the existence of substantial training data from both the source and the target speakers. Furthermore, in some cases, since the training criteria of the conversion function is based on least squared error estimation [2, 4, 5], the training data must be parallel, that is, the source speaker and the target speaker must speak the same sentences for training.

Although the requirement for parallel training data is often acceptable, there are applications which require voice transformation for previously unknown speakers. Examples can be found in the entertainment and media industries where users transform their voices to sound like well-known personalities. Other uses include pronunciation correction in interactive language learning, voice restoration for impaired speakers needing to make telephone calls, and identity masking for telephone services where the user wishes to remain anonymous. In this paper, we present a modified voice conversion system which requires no pre-existing training data from the source speaker. This system is an extension of our previous work [5] which followed the conventional approach of linear transformations trained using least square error criteria on parallel data. In the modified system described here, the target speaker's training data is stored in a database and labelled by aligning it with the states of a HMM-based recogniser. The speech of a previously 'unseen' speaker is then used to index into this database to extract target frames which can be paired with the source in order to estimate the required conversion functions. The remainder of the paper is organised as follows. In section 2, the basic voice conversion scheme is outlined along with the details of the new procedure for handling unknown speakers. In Section 3, a specific implementation of the system is described and an objective distortion metric is used to compare the transformed speech with the target speech. In section 4, a subjective evaluation of the overall performance based on listening tests is described. Finally, our conclusions are presented in section 5.

2. The Voice Conversion System

2.1. Existing System

The voice conversion system for unknown speakers is an extension of the system presented in [5]. In this system, a pitch synchronous sinusoidal model is used for speech signal representation and modification. Speaker identity is represented by a 15 dimensional vector $\mathbf{x}$ of line spectral frequencies (LSF) used to encode the spectral envelope and these are transformed using a set of M interpolated linear transforms W_m such that the target $\mathbf{y}$ is given by:

$$\mathbf{y} = (\sum_{m=1}^{M} \lambda_m(\mathbf{x}) W_m)\bar{\mathbf{x}} \qquad (1)$$

where $\bar{\mathbf{x}} = [\mathbf{x}', 1]'$ is the extended vector of $\mathbf{x}$ and λ_m is the interpolation weight of transform matrix W_m which is determined from the posterior probabilities of an M component Gaussian Mixture Model (GMM) trained on the source data[2].

In order to reduce distortion and unwanted artifacts, this core conversion scheme is in practice augmented by a number of refinements. These include a residual selection scheme, a phase predictor and perceptually-based spectral enhancement. In addition, since most unvoiced sounds have little vocal tract structure and cannot be regarded as short term stationary signals, their spectral envelopes show large variations. Hence, it is not effective to apply linear transformation to convert them. Instead, a unit selection technique is used[5][1]. Since all of these refinements depend only on the target training data, they can easily be carried over to a system for converting the speech of an unknown speaker. Thus, the key problem

[1] Examples of the converted speech generated by this system can be found in http://mi.eng.cam.ac.uk/∼hy216/VoiceMorphingPrj.htm.

to solve is that of estimating the transformation matrices W_m when there is no pre-existing source data.

2.2. Estimating Transformations from Limited Data

2.2.1. Least Square Error Estimation

In the conventional case, where parallel time-aligned source and target data are available, the estimation of the transformation matrices W_m is based on the least square error criteria. For example, a single global matrix W can be estimated by minimizing

$$E = \frac{1}{2}\sum_{t=1}^{N}\Big[\mathbf{y}_t - \mathbf{W}\bar{\mathbf{x}}_t)\Big]^T\Big[\mathbf{y}_t - \mathbf{W}\bar{\mathbf{x}}_t)\Big] \tag{2}$$

where $\mathbf{W}$ is a $p \times (p+1)$ dimensional matrix, and $\mathbf{x}_t$ is the source LSF vector at time t and $\mathbf{y}_t$ is the corresponding target LSF vector. The standard solution to this problem is given by

$$\mathbf{W} = \Big(\sum_{t=1}^{N}\mathbf{y}_t\bar{\mathbf{x}}_t^T\Big)\Big(\sum_{t=1}^{N}\bar{\mathbf{x}}_t\bar{\mathbf{x}}_t^T\Big)^{-1} \tag{3}$$

The case of multiple transforms is a straightforward extension of this (see [2]).

2.2.2. The Limited Source Data Case

When there is no pre-existing source data, the required transformation functions must be estimated using only the source speech to be transformed. In this case, the problem reduces to that of finding one (or more) frames from the target data to associate with each frame of the unknown source. In the system described here, this is achieved by using a speech recognizer to index the target training data so that each unknown source frame can be used to retrieve similar frames from the target database. The retrieved frames are then used to estimate conversion functions in a similar way to conventional methods.

In more detail, the procedure is as follows. Firstly, the training data for each target is organised to form a database. This stage is performed just once before transforming any unknown speech:

1. use a speaker independent HMM-based speech recognizer to force align the target data, then label each target speech frame with a state id such that each utterance can be represented by a state sequence[2];
2. parameterise each frame of the target training data and save each target spectral LSF vector $\mathbf{x}$ and its corresponding state id in a database.

Secondly, for each unknown utterance to be transformed

3. process the unknown utterance with the same HMM-based recogniser used to label the training data. If the orthography of the utterance is known, forced alignment can be used. Otherwise, the utterance must be recognised. The end result is to label each frame of the unknown speech with a HMM state id;
4. using the state id sequence of the unknown source, select corresponding target vectors from the target database using a criteria that encourages longest matching state sequences. This '*unit selection*' step results in a one to one mapping between the input source spectral vectors and the target spectral vectors;
5. based on this mapping, compute a linear transformation as described in 2.2.1 (the estimation of multiple transforms is dealt with later).

The unit selection step in this training approach is rather crucial. In order to ensure that the continuous spectral evolution of the source is carried over into the transformed speech, it is important to select continuous target sequences wherever possible. This is why a criteria that encourages longest matching state sequences is used. An example will illustrate how this algorithm works. If the sequence of source state ids is "1 1 1 2 2 2 2 3 3" , and the longest matching sequence in the target database is "1 1 1 2 2" then the target spectral vectors corresponding to this subsequence are extracted. The procedure then repeats looking for a match for "2 2 3 3" and so on until the whole of the source sequence is matched. The selected target vectors are then concatenated to form the counterpart of the source vectors.

It should perhaps be noted that the number of parallel vectors which can be extracted from the voiced sounds in one utterance is generally not sufficient to train a robust transformation matrix. In this case, however, it is the training data which is to be transformed and by definition of the least squares criterion, the estimated matrix does provide, in some narrow sense at least, an adequate transformation of the training data. Also, it should be noted that applying a global transformation is better than simply copying the target vectors, as the latter results in discontinuities which lead to unnatural output speech. Using a transformation avoids this problem since the converted vectors retain the continuity inherent in the source.

2.3. Multiple Transforms

The basic system described above for transforming the voice of an unknown speaker assumes the use of a single global linear transform. However the use of a single transform results in a significant averaging effect on the formant structure and the use of multiple transforms generally delivers better quality [4].

To use multiple transforms when the source speaker is unknown, we use the same interpolation formula as equation 1, however, the GMM and corresponding mixture weights λ_m are estimated using the target data rather than the source. This is done in two passes. In the first pass, a single global transform is estimated and applied to convert the source vectors $\mathbf{x}$. In the second pass, the posterior probabilities of the target GMM components given the converted vectors $\tilde{\mathbf{x}}$ are computed; these posterior probabilities are then used as the interpolation weights.

After obtaining the interpolation weights, the estimation of multiple transforms can then in principle proceed as in the conventional parallel data case. However, in practice, the shortage of source training data makes it difficult to robustly estimate multiple transforms. For example, there will be 400 voiced frames in a typical 4 seconds long utterance. This is just sufficient to train a single 15×16 dimensional linear transform matrix but not more. We solve this problem by applying a more aggressive unit selection approach whereby each frame of the unknown source is matched with multiple frames from the target database. More specifically, the unit selection process is performed iteratively until enough vectors (e.g. 300 × number of GMM components) have been extracted from the target database. This means that each source vector appears several times in the training data, each time paired with a different target vector.

3. Implementation and Objective Testing

3.1. Target Database and Test Data

To build a system for evaluation, each target speaker recorded about 650 reading TIMIT sentences. These sentences were then force aligned using a set of speaker independent monophone HMMs trained using the HTK Toolkit on the WSJ0 British English read speech database. Note that the speech features used by the HMMs

[2] We assume that there is enough target speaker data to cover all HMM states.

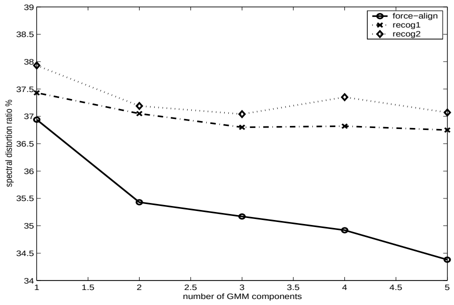

Figure 1: *Spectral distortion ratio over different numbers of GMM components. force-align: voice conversion using force alignment; recog1: voice conversion using LVCSR recogniser (86.4% phones correct); recog2: voice conversion using monophone recogniser (55.2% phones correct).*

are MFCCs whereas LSFs are used for the voice conversion.[3] Given the forced alignment information, each speech frame was then labelled with a HMM state id, such that each target utterance can be represented by a state id sequence. At the same time, a 15 dimensional LSF vector was computed for each speech frame. The state sequences and their corresponding LSF vectors were then stored in a database.

The VOICES database from OGI [3] was used for evaluation. This corpus contains recorded speech from 12 different speakers reading 50 phonetically rich sentences none of which appear in the target speech database. For comparison, each target speaker also recorded a version of these OGI test sentences.

3.2. Objective Distortion Measure

A log spectral distortion metric was used to provide an objective measure of conversion performance

$$d(S_1, S_2) = \sum_{k=1}^{L} (loga_k^1 - loga_k^2)^2 \tag{4}$$

where $\{a_k\}$ are the amplitudes resampled from the spectrum S at L uniformly spread frequencies. A distortion ratio is then used to compare the converted-to-target distortion with the source-to-target distortion, which is defined as,

$$D = \frac{\sum_{t=1}^{N} d(S_{tgt}(t), S_{cov}(t))}{\sum_{t=1}^{N} d(S_{tgt}(t), S_{src}(t))} \times 100\% \tag{5}$$

where $S_{tgt}(t)$,$S_{src}(t)$ and $S_{cov}(t)$ are the target spectrum, source spectrum and the converted spectrum at time t respectively. N is the total number of test vectors.

3.3. Transcription Mode and Number of Transforms

If the orthography of the unknown speech is available, forced alignment can be used and phone recognition errors avoided. When there is no transcription of the source, some form of recognition must be used. Experiments were therefore conducted to quantify the cost incurred by using a recogniser instead of forced alignment. At the same time, the effects of using single versus multiple transforms were recorded.

Two types of recogniser were compared: a one-pass HTK large vocabulary recogniser (recog1) [6] yielding 86.4% phones correct and a simple monophone recogniser (recog2)yielding 55.2% phones correct. The same monophones were also used for the forced-alignments.

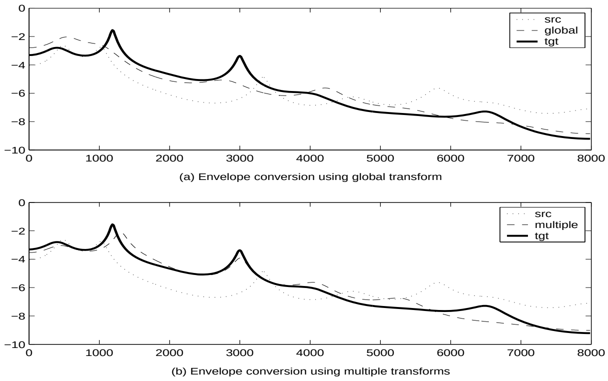

Figure 2: *Spectral envelope conversion using (a) global transform and (b) multiple transforms.*

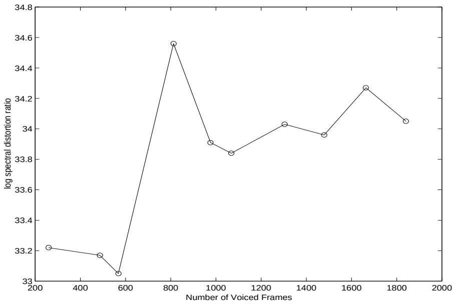

Figure 3: *log spectral distortion over different number of training voiced frames.*

As shown in Fig. 1, the spectral distortion ratio of the systems using recognition are always higher than that of the system using forced alignment, and the difference increases when using multiple transforms where mis-recognition errors on the voiced sounds result in significant distortion. The large vocabulary recogniser is consistently better than the phone recogniser. Indeed, the performance of the latter does not improve monotonically with increasing numbers of transforms. Presumably this is a consequence of the higher phone error rate. An example of spectral envelope conversion using a global transform and multiple transforms in the forced alignment case is displayed in Fig. 2. This figure clearly demonstrates that the spectral envelope converted by multiple transforms matches the target envelope more closely than that of the global transforms.

3.4. Training Data Size vs Spectral Distortion

In the previous experiments, the linear transforms were trained based on a single user utterance which normally is just a few seconds long. To determine the effect of increasing the number of voiced frames in the source training data, an experiment was conducted to convert source utterances with total voiced speech ranging from 2 to 20 seconds. In each case, a fixed number of 8 linear transforms were used. The results are shown in Fig. 3 where it can be seen that for very short segments, increasing the number of voiced frames from around 250 to 600 results in a decrease in spectral distortion. However, after that the log spectral distortion seems to have no clear correlation with the amount of training data. The probable reason for this is that increasing the training data size increases the spectral variability in the source at a similar or greater rate than the improvement in the transforms' ability to model it. The net result overall is increased distortion.

4. Subjective Evaluation

Subjective evaluations were also carried out to assess the overall performance of this voice conversion system and provide contrasts

[3] Preliminary experiments indicated that using LSFs for speech recognition significantly degraded performance

for the performance of forced alignment vs recognition, and global transform vs multiple transforms. Specifically, two kinds of formal listening tests have been performed; the standard ABX test and a similarity test.

4.1. ABX test

In the ABX test, a panel of listeners were asked to judge whether an utterance X sounded closer to utterance A or B in terms of speaker identity, where X was the converted speech and A and B were either the source speech or the target speech. The converted speech consisted of utterances generated by four methods, 1) forced alignment plus a global transform (FG); 2) forced alignment plus multiple transforms (FM); 3) recognition plus a global transform (RG); and 4) recognition plus multiple transforms (RM).

Table 1 gives the percentage of the converted utterances that were labelled as closer to the target in the above four categories. As can be seen, the use of multiple transforms gives an improvement over using a global transform in the forced alignment case, but degrades the performance in the recognition case. This is consistent with the results obtained using the objective distortion measure. Since a global transform represents an average across all phones, the effects of recognition errors are smoothed out. On the other hand, using multiple transforms enables context dependent effects to be modelled and in the absence of phone errors, the results are significantly better.

Table 1: *Results from the ABX test.*

FG	FM	RG	RM
81.7%	88.3%	76.7%	68.3%

4.2. Similarity test

In an effort to further assess the overall performance of the system, a similarity test was designed in addition to the ABX test. In this test, pairs of utterances were presented to the listeners. Listeners were asked to rate the similarity of each pair in terms of speaker identity on a scale between 10 for “identical” to 0 for “totally different”. These pairs of utterances cover the combinations among target speaker (T), source speaker (S), converted speaker using force alignment plus multiple transforms (F), converted speaker using recognition plus global transform (G), converted speaker using recognition plus multiple transforms (R), and converted speaker using forced alignment plus multiple transforms plus target prosody (P). In the ‘P’ case, the prosody of the source utterance was altered to match as closely as possible the prosody of the corresponding target utterance. Its purpose was to evaluate the influence of prosody on the perception of speaker identity. Different sentences were used to make these pairs so that the listeners would not be distracted by the sentence level similarity, but would make judgements based only on speaker identity.

Fig. 4 presents the results of this similarity test. For each of the listeners the mean scores of each of the cases were computed, and the variations of the mean scores over the panel of listeners are displayed as a vertical line. The means of all the scores in each case are marked by an “X”. It can be seen from this figure that the converted speaker sounds very different from the source speaker as the “FS” score is very low, whilst the “FT” score indicates that the source speaker has been converted to be quite close to the target speaker. Moreover, if the source prosody can be transformed to match the target prosody, the converted speech sounds very similar to the target speaker.

5. Conclusion

This paper has described a method of converting the speech of an unknown speaker to sound like that of some designated target speaker. The approach utilises speech recognition technology to generate a mapping between the source spectral features and target spectral features extracted from a database. This then enables the least square criteria to be applied to estimate linear transformations. Furthermore, by using an aggressive target database retrieval strategy which matches each source vector with multiple target vectors, multiple conversion transforms can be estimated for quite short source utterances.

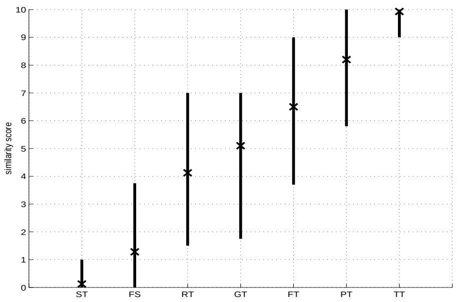

Figure 4: *Similarity Test. The symbol ‘S’ denotes source speaker; ‘T’ target speaker; ‘F’ converted speaker using forced alignment and multiple transforms; ‘R’ converted speaker using recognition and multiple transforms; ‘G’ converted speaker using recognition and a global transform; and ‘P’ converted speaker using forced alignment and multiple transforms plus target prosody.*

The effectiveness of the system has been investigated using both objective and subjective measures, and the results from each are broadly consistent. Firstly, it has been shown that the use of a single global transform is sufficient to hide the identity of the source speaker and provide a voice which is somewhat like the target. However, for a close similarity to the target, multiple transforms are required. In the system described, the latter can be robustly trained but only when a transcription of the source utterance is available so that forced alignment can be used. When recognition is used, phone errors result in inappropriate transformations being learnt and reduced performance results. Thus, further work is needed to find methods of reducing the effects of phone errors. perhaps, by making use of confidence measures. Finally, the similarity tests also demonstrated the need for good prosody transformation to complement spectral transformation and this is clearly another area needing further work.

6. ACKNOWLEDGMENTS

This work was supported by a grant from Anthropics Technology Ltd. The authors thank the volunteers of the perceptual tests for their assistance.

7. References

[1] Abe, M., Nakamura, S., Shikano, K. and Kuwabara, H., “Voice conversion through vector quantization”, ICASSP, 1988.

[2] Ye, H. and Young, S.. “Perceptually Weighted Linear Transformation for Voice Conversion”, Eurospeech 2003.

[3] Kain, A.. “High resolution voice transformation”, PhD dissertation, OGI, 2001.

[4] Stylianou, Y., Cappe, O. and Moulines, E., “Continuous probabilistic transform for voice conversion”, IEEE Trans. Speech & Audio Processing, vol. 6, pp. 131-142, 1998.

[5] Ye, H. and Young, S.. “High Quality Voice Morphing”, in Proc. of ICASSP, 2004.

[6] Odell, J., Valtchev, V., Woodland, P.C. and Young, S.J.. “A One Pass Decoder Design for Large Vocabulary Recognition,” Proceedings of Human Language Technology Workshop, pp. 405-410, Mar. 1994.