Novel Method for Data Clustering and Mode Selection with Application in Voice Conversion

Jani Nurminen, Jilei Tian and Victor Popa

Multimedia Technologies laboratory
Nokia Research Center, Tampere, Finland
{jani.k.nurminen, jilei.tian, ext-victor.popa}@nokia.com

Abstract

Since the statistical properties of speech signals are variable and depend heavily on the content, it is hard to design speech processing techniques that would perform well on all inputs. For example, in voice conversion, where the aim is to transform the speech uttered by a source speaker to sound as if it was spoken by a target speaker, different types of inter-speaker relationships can be found from different types of speech segments. To tackle this problem in a robust manner, we have developed a novel scheme for data clustering and mode selection. When applied in the voice conversion application, the main idea of the proposed approach is to first cluster the target data to achieve a minimized intra-cluster variability. Then, a mode selector or a classifier is trained on aligned source-related data to recognize the target-based clusters. Auxiliary speech features can be used to enhance the classification accuracy, in addition to the source data. Finally, a separate conversion scheme is trained and used for each cluster. The proposed scheme is fully data-driven and it avoids the need to use heuristic solutions. The superior performance of the proposed scheme has been verified in a practical voice conversion system.

Index Terms: voice conversion, data-driven, clustering

1. Introduction

One of the common challenges in several areas of speech processing research is caused by the complexity of speech signals. The signals are generated through complicated speech production mechanisms and consequently the signals possess highly variable statistical properties. Moreover, each person has her or his own unique physical properties related to speech production. As a result, it is very challenging to develop speech processing techniques that would perform robustly and well on all input speech signals. For example, nearly stationary voiced regions should usually be treated differently than plosives. Numerous different approaches have been proposed to tackle this problem in different areas of speech processing research, usually based on some heuristic solutions. Nevertheless, no universal solution exists for the problem.

Voice conversion is one of the speech related research topics in which the signal processing techniques have to operate with different kinds of speech signals and speech signal regions. In voice conversion, the goal is to convert the speech signal from a source speaker to sound as if it was uttered by a target speaker, without changing the actual speech content. Typical approaches for voice conversion include Gaussian mixture modeling (GMM) based conversion [1], neural network based conversion [2] [3], hidden Markov model (HMM) based conversion [4], linear transformation based conversion [5] [6], and codebook based conversion [7]. For the particular problem of handling different types of speech segments, the solutions proposed in the literature include the use of acoustic similarity based classification and regression trees [8], phoneme-tied codebooks [9], K-means based clustering [10], and phoneme-based modeling [11].

In this paper, we propose a novel solution for handling different types of speech segments in a completely data-driven way using different modes. When applied in voice conversion, the proposed approach for data clustering and mode selection is based on the idea that for the training of the multiple processing schemes the data is split into different clusters using the best possible clustering on the target data. In this way, the intra-cluster behavior of the data becomes easy to model. For the mode selection, a different approach has to be used since the target data itself is not available. The solution proposed in this paper is to train a classifier that aims to recognize the correct target based cluster using only source-related data features.

This paper is organized as follows. Section 2 introduces the voice conversion system that served as a target application for the proposed technique. The novel method for data clustering and mode selection is described in Section 3. In Section 4, we demonstrate the very good performance of the method using practical test results. Finally, Section 5 concludes the paper with some summarizing remarks.

2. Application: a voice conversion system

The development of the proposed method for data clustering and mode selection was motivated by the fact that it is very hard to build a conversion model that could handle well all kinds of input data. Since our voice conversion system is also used as a test platform in Section 4 of this paper, the system is described at a general level in the rest of this section. For more detailed information about the voice conversion system, please refer to [12].

The voice conversion system that served as a target application is based on parametric modeling of speech. The speech model separates the speech signal into the vocal tract contribution represented as line spectral frequencies (LSFs) and into a parametric excitation model. The excitation signal is represented using an approach based on sinusoids and noise. The excitation related parameters consist of the pitch

parameter, the energy, the spectral excitation amplitudes, and of voicing information for the spectrum.

The conversion of the speech parameters is generally handled one parameter at a time, using a GMM based approach. Let **x** and **y** denote parameter vectors associated with the source and the target speakers, respectively. For the training of a conversion model, combined source-target feature vectors are generated by joining aligned source and target vectors, denoted as $\mathbf{z}=[\mathbf{x}^T\ \mathbf{y}^T]^T$, that can be used to train a conversion model. In the training, we have used the popular approach proposed in [13] that makes use of the aligned data **z** to estimate the GMM parameters (α, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) of the joint density p(**x**,**y**). This is accomplished iteratively through the well-known Expectation Maximization (EM) algorithm [14]. The conversion of the speech parameters follows a scheme where the trained GMM parameterizes a piece-wise linear mapping function that minimizes the mean squared error (MSE) between the converted source and target vectors. This conversion function is constructed as shown in [13]:

$$F(\mathbf{x}) = E(\mathbf{y} \mid \mathbf{x}) = \sum_{i=1}^{L} p_i(\mathbf{x}) \cdot \left(\boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx}\left(\boldsymbol{\Sigma}_i^{xx}\right)^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_i^x\right)\right), \quad (1)$$

where

$$p_i(\mathbf{x}) = \frac{\alpha_i \cdot N(\mathbf{x}, \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx})}{\sum_{j=1}^{L} \alpha_j \cdot N(\mathbf{x}, \boldsymbol{\mu}_j^x, \boldsymbol{\Sigma}_j^{xx})}. \quad (2)$$

The covariance matrix of the *i*-th Guassian mixture is constructed as

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy} \\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy} \end{bmatrix}. \quad (3)$$

Similarly,

$$\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_i^x \\ \boldsymbol{\mu}_i^y \end{bmatrix}, \quad (4)$$

represents the mean vector of the *i*-th Gaussian mixture of the GMM.

3. Proposed approach for data clustering and mode selection

The proposed approach for data clustering and mode selection starts from the idea that the data should preferably be clustered into different operating modes in the best possible manner from the viewpoint of effective processing. Often, this best possible clustering is based on data that is not yet available during usage. For example, in the case of voice conversion, if the target is to minimize the potential conversion error, the most effective approach would be to cluster the combined data from the source and the target side into different processing modes based on the target data. This choice ensures a minimized potential conversion error within each mode or cluster. However, the corresponding mode selections would not be possible during usage since the target data is not available at that time. Nevertheless, our proposal is that the training data is initially clustered using the most optimal clustering approach. Then, the next step is to train a mode selector that aims at finding the correct cluster based on the data that is available during usage. This data can include in addition to the conventionally available data any auxiliary features that can be made available. Finally, a separate processing scheme is trained and used for each mode.

When applied in the voice conversion task introduced in Section 2, the proposed approach first finds *M* clusters solely based on the target data features **y**. For example, if the aim is to convert vectors containing line spectral frequency data, the initial clustering is performed based on target LSF vectors only. The clustering can be performed e.g. using the well-known K-means algorithm to obtain the clusters $\mathbf{y}^{(1)}$, $\mathbf{y}^{(2)}$, …, $\mathbf{y}^{(M)}$.

After obtaining the initial grouping, the next step is to train a mode selector with the aim to recognize the target based clusters using only data from the source speaker. To facilitate the classification task, auxiliary features derived from the source data can be used in addition to the source vectors **x**. In principle, this auxiliary data denoted as **aux** can include any/all the features that one can *extract from the source data.* For example, the auxiliary feature set could include acoustic parameters such as pitch, voicing and energy as well as other parameters such as phoneme information, linguistic location, linguistic duration and part-of-speech. Given the initial target-based clusters, the extended aligned data set, denoted now as $\mathbf{z} = [\mathbf{x}^T\ \mathbf{aux}^T\ \mathbf{y}^T]^T$, can be straightforwardly split into the same *M* groups, $\mathbf{z}^{(1)}$, $\mathbf{z}^{(2)}$, …, $\mathbf{z}^{(M)}$. Based on this grouping, it is possible to train a classifier aiming to find the correct cluster using only the source related data vector $[\mathbf{x}^T\ \mathbf{aux}^T]^T$. We implemented the classifier using a simple linear discriminative function $\mathbf{D}\cdot[\mathbf{x}^T\ \mathbf{aux}^T]^T$ but it would also be possible to use other techniques such as non-linear discriminative functions, neural networks or support vector machines. The exact selection of the auxiliary feature set is not a highly critical issue in the sense that the features with no additional discriminative information will receive a very low or even a zero weight in the training while the more relevant features will receive a larger weight.

Once the mode selector or the classifier is available, a separate conversion scheme is trained for every mode with a training data set belonging to that mode. It is possible to either use the training data sets based on the initial clustering that was made based only on the target data or to re-cluster the data using the trained classifier to obtain re-grouped training data sets. The latter approach provides enhanced robustness against classification errors, and thus it should preferably be followed in cases where the classification error rate is not very low.

During the usage of the multi-mode processing system, the conversion system must first obtain the source vector to be converted and the corresponding auxiliary vector. This data is used as an input to the classifier that selects the mode. Finally, the conversion of the vector is handled using the conversion scheme corresponding to the selected mode.

The proposed approach summarized in Figure 1 and Figure 2 has many beneficial properties. First, the approach is fully data-driven and there is no need to rely on any heuristic solutions. Second, the method is very flexible in the sense that it is for example very easy to change the number of modes/clusters. Third, there is no requirement to utilize any linguistic information but if such information is available it can very easily be used to support the mode selection. Finally,

Training Algorithm:
Step 1: Define and extract an auxiliary feature set **aux** from the source training data set;
Step 2: Align the source related data and the target data to form extended combined feature vectors $\mathbf{z} = [\mathbf{x}^T \mathbf{aux}^T \mathbf{y}^T]^T$;
Step 3: Split the target data **y** into *M* clusters using e.g. the K-means algorithm;
Step 4: Group the extended vectors **z** into the same *M* clusters based on the clustered target data **y**;
Step 5: Train a mode classifier that aims at finding the correct target based cluster using only the source related features **x** and **aux**;
Step 6: Train *M* separate models for the different modes. Use as the training data the data classified to the corresponding cluster;

Figure 1. *Training algorithm for the proposed approach.*

Conversion Algorithm:
Step 1: Extract the auxiliary feature vector **aux** from the source data;
Step 2: Select the correct mode using the source related vectors **aux** and **x** as input;
Step 3: Use the selected model to convert the source feature vector **x**;

Figure 2. *Conversion algorithm for the proposed scheme.*

the proposed method offers very good performance. The performance advantage is demonstrated in Section 4 using practical experiments but the good performance can also be explained from another point of view. Figure 3 depicts the distribution of the first two LSFs in a small set of target LSF vectors, selected randomly from a larger voice conversion training set. The line illustrates the boundary between the ideal clusters in the case of two clusters, whereas the circles and crosses demonstrate the clustering decisions based on the widely used voiced/unvoiced classification. Provided that the mode selection is made correctly, it is evident that in the case of optimal clustering the distribution of any conversion errors will be much narrower than in the case of voiced/unvoiced clustering. While it is in general not possible to achieve a 100% mode classification rate, the proposed approach still successfully mimics this optimal case, leading to clear measurable improvements.

4. Experimental results

The proposed data clustering and mode selection approach was tested in the voice conversion system introduced in Section 2. To highlight the performance advantage achievable using the novel clustering method, we compared it against the common approach of using voiced/unvoiced clustering that also offers good performance. The comparison was done through the measurement of the average mean squared error between the converted and the target vectors.

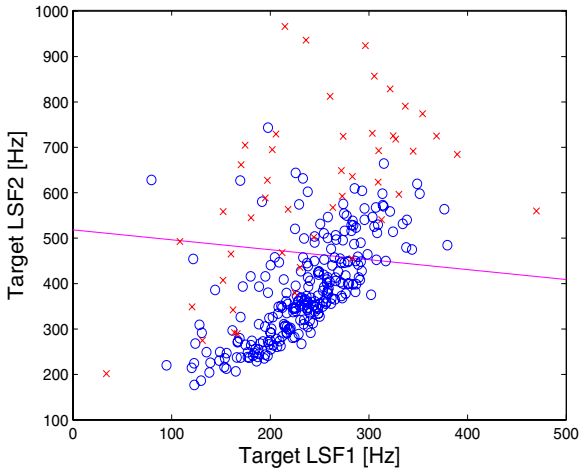

Figure 3. *Ideal clustering vs. voiced/unvoiced clustering. The line illustrates the division between the two ideal clusters while o and x denote voiced and unvoiced data, respectively. It is easy to see that there is significantly less variability within each cluster in the case of ideal clustering.*

4.1. Test set-up

The two different conversion schemes, the first based on the proposed approach and the second based on the traditional voiced/unvoiced clustering, were implemented for the conversion of LSF vectors. In the implementation of the proposed approach, we used several source-speech related features to form the auxiliary data vector **aux**. More specifically, the auxiliary data included the first and second derivatives of the LSF vectors (dLSF/dt and d_2LSF/dt^2), the pitch parameter, the energy parameter, the residual amplitude spectrum and the voicing information for the spectrum. The mode selector was implemented using a simple linear discriminative function. In the case of the traditional voiced/unvoiced clustering, we made a single voicing decision for each frame based on the voicing information for the spectrum.

Both conversion schemes were trained and tested using the same training and testing data sets. A data set containing 90 sentences (29 880 frames) from a source speaker and a target speaker was used for the training while a distinct set of 99 sentences (32 700 frames) was reserved for the testing phase. In both sets, the source and target vectors were aligned using dynamic time warping supported with phoneme-level segmentation. All the conversions were handled using the Gaussian mixture modeling based approach summarized in Section 2. One 16-mixture GMM was trained for each mode.

Since the mode classifier was implemented in a very simple way, we also implemented a third conversion scheme that directly utilized the perfect clustering based on the target data. In general, of course, the implementation of such a perfect classifier is not possible. Nevertheless, this third conversion scheme can be used for demonstrating the theoretical performance bound that cannot be exceeded with the proposed approach provided that the initial clustering and the conversion schemes are kept unchanged.

4.2. Results

The results achieved in the test are summarized in Table 1. For the scheme based on the conventional voiced/unvoiced clustering, the mean squared error between the converted LSFs and the corresponding target LSFs was 23058 for the training set and 23559 for the testing set. For the proposed scheme, when implemented as described above, we achieved the MSE scores of 21015 and 21833 for the training set and the testing set, respectively. Since our classifier was implemented in a very simple way, we also tested the performance in the ideal hypothetical case with 100% classification rate. In this ideal case, providing the performance bound for the given initial clusters, the MSE figures were found to be 15295 and 15770 for the training and the testing set, respectively.

As is clearly evident from the results, the proposed method outperforms the conventional approach with a clear margin, despite the fact that the simple mode classifier only achieved a classification error rate of 12.4%. Moreover, the performance advantage was achieved even though the traditional voiced/unvoiced classification used as a reference also offered a very natural and efficient clustering scheme. For example, we have found in our earlier experiments that this voiced/unvoiced scheme already clearly outperforms an implementation with only one GMM model but twice the number of mixtures. If the proposed approach was compared against some arbitrary clustering scheme, the improvement would have most likely been even larger.

Table 1. *Comparison between the conversion MSE achieved using the conventional voiced/unvoiced clustering and the proposed data-driven clustering schemes.*

	Training Set	Testing Set
Voiced/unvoiced clustering	23058	23559
Proposed	21015	21833
Proposed (perfect classifier)	15295	15770

5. Conclusions

This paper has described a novel approach for data clustering and mode selection. The main idea is to first perform the clustering in the optimal manner and then to train a mode selector that aims to find the correct cluster based on the data that is available during the usage time. Finally, a separate processing scheme is trained for each of the clusters/modes. The proposed techniques have been implemented in a voice conversion system, and the very good performance has been verified in practical experiments. In addition to the performance advantage, the proposed method is very flexible and it enjoys the benefit of being a completely data-driven technique, eliminating the need to use any heuristic solutions or linguistic knowledge.

6. Acknowledgements

This work has partially been funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech-to-Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org).

7. References

[1] Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H., "Voice conversion through vector quantization", in Proc. International Conference on Acoustics, Speech, and Signal Processing, New York, NY, USA, pp. 655-658, 1988.

[2] Narendranath, M., Murthy, H., Rajendran, S., and Yegnanarayana, B., "Transformation of formants for voice conversion using artificial neural networks", Speech Communication vol.16, pp. 207-216, 1995.

[3] Watanabe, T., Murakami, T., Namba, M., Hoya, T., and Ishida, Y., "Transformation of spectral envelope for voice conversion based on radial basis function networks", in Proc. International Conference on Spoken Language Processing, Denver, USA, 2002.

[4] Kim, E.-K., Lee, S., and Oh, Y.-H., "Hidden Markov Model based voice conversion using dynamic characteristics of speaker", in Proc. European Conference on Speech Communication and Technology, Rhodes, Greece, 1997.

[5] Ye, H. and Young, S., "Perceptually weighted linear transformations for voice conversion" in Proc. of Eurospeech'03, Geneva, Switzerland, 2003.

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

[7] Arslan, L. M. and Talkin, D., "Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum", in Proc. Eurospeech'97, Rhodes, Greece, pp. 1347-1350, 1997.

[8] Duxans, H., Bonafonte, A., Kain, A., van Santen, J., "Including dynamic and phonetic information in voice conversion systems", in Proc. International Conference on Spoken Language Processing, Jeju Island, Korea, 2004.

[9] Kang, Y., Shuang, Z., Tao, J., Zhang, W., and Xu, B., "A Hybrid GMM and Codebook Mapping Method for Spectral Conversion", in Proc. ACII 2005, pp.303-310, 2005.

[10] Suendermann, D., Bonafonte, A., Ney, H., and Hoege, H., "Voice Conversion Using Exclusively Unaligned Training Data", In Proc. ACL/EMNLP 2004, Barcelona, Spain, 2004.

[11] Kumar, A. and Verma, A., "Using phone and diphone based acoustic models for voice conversion: a step towards creating voice fonts" in Proc. International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, 2003.

[12] Nurminen, J., Popa, V., Tian, J., and Kiss, I., "A parametric approach for voice conversion", to be published in the TC-STAR workshop on Speech-to-Speech Translation, Barcelona, Spain, 2006.

[13] Kain, A. and Macon, M.W., "Spectral voice conversion for text-to-speech synthesis", in Proc. ICASSP'98, Seattle, WA, USA, pp. 285-288, 1998.

[14] Dempster, A.P., Laird, N.M., and Rubin, D.B., "Maximum likelihood from incomplete data via the EM algorithm", Journal of Royal Statistical Society B vol.39, pp. 1-38, 1977.