

A Novel Technique for Voice Conversion Based on Style and Content Decomposition with Bilinear Models

Victor Popa¹, Jani Nurminen², Moncef Gabbouj¹

¹Department of Signal Processing, Tampere University of Technology, Tampere, Finland

²Devices, Nokia, Tampere, Finland

victor.popa@tut.fi, jani.k.nurminen@nokia.com, moncef.gabbouj@tut.fi

Abstract

This paper presents a novel technique for voice conversion by solving a two-factor task using bilinear models. The spectral content of the speech represented as line spectral frequencies is separated into so-called style and content parameterizations using a framework proposed in [1]. This formulation of the voice conversion problem in terms of style and content offers a flexible representation of factor interactions and facilitates the use of efficient training algorithms based on singular value decomposition and expectation maximization. Promising results in a comparison with the traditional Gaussian mixture model based method indicate increased robustness with small training sets.

1. Introduction

Voice conversion technology enables to transform one speaker's speech pattern into speech having distinct identity characteristics of another speaker, while preserving the original content or meaning. Though commercial usage of the voice conversion techniques has not been very popular yet, the interest has risen immensely over the last few years. One of the reasons is the attractive idea to use voice conversion in cost-effective personalization of TTS systems. Without voice conversion, new voices have to be created in a time-consuming and expensive way using extensive recordings and manual annotations. Voice conversion can also be used to make a synthetic voice speak in languages that the original voice talent cannot speak. Other applications for voice conversion include security related usage to hide the identity of the speaker and entertainment applications, etc.

The research on voice conversion has received an increasing amount of attention, and several different voice conversion approaches have been proposed in the literature. From the technical point of view, typical approaches presented in the literature include Gaussian mixture modeling (GMM) [2], frequency warping [3], artificial neural networks [4], hidden Markov models (HMM) [5], linear transformation [6] and codebook based conversion [7]. Despite the large number of voice conversion techniques, the problem has not been fully solved yet and all of the current techniques have their weaknesses. For example GMM based voice conversion can convert the speaker identity relatively well, but the converted speech usually has compromised quality due to over-smoothing, formant broadening, etc. Similarly it is known that while the quality degradation caused by frequency warping based conversion method is very small the identity conversion is not as successful because the frequency warping fails to properly map the formant amplitudes and bandwidths.

Perceptual systems of living organisms are capable of separating "content" and "style" factors that underlie a set of observations. Starting from this need of processing two

independent factors separately a framework for solving this kind of two-factor tasks was introduced in [1]. By fitting simple and tractable bilinear models to a training set of observations, the underlying style and content factors can be efficiently separated in a flexible representation that facilitates generalization to unseen styles or content classes.

This paper argues that speech fits naturally to this perceptual framework and can be modeled as an interaction between styles (different speakers) and content (actual spectral/phonetic content). Starting from this idea we have developed a novel technique for voice conversion that in theory also allows very efficient speech compression. In this paper, the proposed voice conversion approach is applied to convert the spectral content modeled by line spectral frequencies (LSF). The experiments show that the converted speech outperforms the traditional GMM method in terms of speaker identity and quality and is significantly more robust when used on small training sets.

The paper is organized as follows. In the next section, we introduce the proposed method based on asymmetric bilinear models. In Section 3, we demonstrate the benefits of the proposed approach through objective measurements and listening test results. Finally, concluding remarks, including some potential future research directions, are presented in Section 4.

2. Voice conversion with asymmetric bilinear models

We briefly introduce in this section the general framework originally presented in [1], insisting on the asymmetric bilinear models and discussing it from the voice conversion perspective. In the following, we will use the terms *style* and *content* to refer to the *voice identity/characteristics* and *spectral/phonetic content*, respectively, which constitute the two independent factors underlying our observations (LSFs).

2.1. Asymmetric bilinear models

In a symmetric model the style s (the speaker's voice characteristics) and content c (the spectral/phonetic content) are represented as parameter vectors denoted a^s and b^c having dimensionalities I and J respectively. Let y^{sc} be an observation vector in style s and content class c and let K be its dimension. In our case y^{sc} is an LSF vector of one speaker and representing a particular spectral/phonetic content (e.g. from a certain speech unit). We represent y^{sc} as a bilinear function of a^s and b^c , in its most general form given by

$$y_k^{sc} = \sum_{i=1}^I \sum_{j=1}^J w_{ijk} a_i^s b_j^c \quad (1)$$

In this formula i , j and k denote elements of the style, content and observation vectors. The terms w_{ijk} describe the

interaction between the content and style factors and are independent of both style and content.

Asymmetric bilinear models are derived from the symmetric bilinear models by allowing the interaction terms w_{ijk} to vary with the style leading to a more flexible style description. The equation (1) becomes

$$y_k^{sc} = \sum_{i,j} w_{ijk}^s a_i^s b_j^c \quad (2)$$

Combining the style-specific terms in equation (2) into

$$a_{jk}^s = \sum_i w_{ijk}^s a_i^s \quad (3)$$

gives

$$y_k^{sc} = \sum_j a_{jk}^s b_j^c \quad (4)$$

Denoting by A^s the $K \times J$ matrix with entries a_{jk}^s we can rewrite the equation (4) as

$$y^{sc} = A^s b^c \quad (5)$$

In this formulation the a_{jk}^s terms can be interpreted as a style specific linear map from the content space to the observation space.

2.2. Model fitting procedure

The objective of the model fitting is to train the parameters of the asymmetric model to minimize the total squared error over the training dataset. This is equivalent to maximum likelihood (ML) [8] estimation of the style and content parameters based on the training data, with the assumption that the data was produced by the models plus independently and identically distributed (i.i.d.) Gaussian noise.

Now, let $y(t)$ denote the t^{th} training observation ($t=1, \dots, T$). In the voice conversion case, $y(t)$ are LSF vectors representing the spectral content at a particular time instant in the speech uttered by a certain speaker (style). The indicator variable $h^{sc}(t)$ takes the value 1 if $y(t)$ is in style s and content class c and 0 otherwise. The total squared error E of the asymmetric model given in equation (5) is computed over the training set using

$$E = \sum_{t=1}^T \sum_{s=1}^S \sum_{c=1}^C h^{sc}(t) \|y(t) - A^s b^c\|^2 \quad (6)$$

In the voice conversion context we can consider e.g. that our training material consists of T_1 LSF vectors of speech uttered by a speaker s_1 (in style $s=s_1$) and T_2 LSF vectors of speech uttered by a speaker s_2 (in style $s=s_2$) and T_3 LSF vectors of speech uttered by a speaker s_3 (in style $s=s_3$). Therefore in equation (6) we would have

$$T = T_1 + T_2 + T_3 \quad (7)$$

Furthermore, each LSF vector falls into one of C content classes which could in principle correspond to phonetically justified units. In the case of parallel training data, the utterances from the three speakers can be time aligned. In this case we can simplify things by assuming a different class for each of the aligned LSFs, thus there will be only one LSF vector from each speaker (style) falling into each content class.

If the training set contains equal number of observations in each style and in each content class there exists a closed form

procedure to fit the asymmetric model using singular value decomposition (SVD).

In the proposed case of parallel and aligned training data, in order to work with standard matrix algorithms, we stack the SC LSFs (K dimensional column) vectors into a single $SK \times C$ matrix

$$Y = \begin{bmatrix} y^{11} & \dots & y^{1C} \\ \dots & \dots & \dots \\ y^{S1} & \dots & y^{SC} \end{bmatrix} \quad (8)$$

We can express now the asymmetric model in compact matrix form

$$Y = AB, \quad (9)$$

where the $(SK) \times J$ matrix A and the $J \times C$ matrix B represent the stacked style and content parameters, respectively

$$A = \begin{bmatrix} A^1 \\ \dots \\ A^s \end{bmatrix}, \quad (10)$$

$$B = [b^1 \quad \dots \quad b^c]. \quad (11)$$

To find the optimal style and content parameters for equation (9) in a least square sense we compute the SVD of $Y=USV^T$. (S is considered to have the diagonal eigenvalues in decreasing order). By definition we choose the style parameter matrix A to be the first J columns of US and we choose the content parameter matrix B to be the first J rows of V^T . There are many ways to choose the model dimensionality J e.g. from prior knowledge, by requiring a desired level of approximation of data, or by identifying an ‘‘elbow’’ in the singular value spectrum.

2.3. Application in voice conversion

One of the tasks that fall under the framework proposed in [1] and which is of particular interest in voice conversion is the *extrapolation* which is illustrated in Table 1.

Table 1: *The extrapolation task illustrated for a set of characters*

A	B	C	D	E
A	B	C	D	E
?	?	C	D	E

Extrapolation refers to the ability to produce equivalent content in a new style, in our case to produce speech as that uttered by a source speaker but with a target speaker’s voice, that is in a new style. Therefore voice conversion is a direct analogy of the extrapolation task. Extending a little bit the concept of voice conversion we can also define it as the generation of speech with a target voice repeating the speech content as it exists in a test sentence available in one or more instances as uttered by one or more source speakers independently.

We can formulate the voice conversion problem as an extrapolation task as follows. Given a training set of parallel speech data from S ‘source’ speakers plus the target speaker, the task is to generate any test sentence in the target voice starting from S utterances of the test sentence corresponding to the S source voices (styles).

In the first phase we need to align the data. Both the training data and the utterances of the test sentence are aligned

for all speakers. Preferably the alignment respects the prosody of one of the S source speakers which will be considered the main source speaker. A so-called *complete* data is formed by combining the source part of the aligned training data and the aligned utterances of the test sentence. In other words the complete data is the concatenation of training and test data of the S source speakers.

During training we fit the asymmetric bilinear model of equation (9) to the S source styles using the closed-form SVD procedure described in Section 2.2. In addition, we also assume that there are as many classes as LSF vectors in the *complete* data (per speaker). This yields a $K \times J$ matrix A^s representing each style (voice) s and a J dimensional vector b^c representing each LSF class c in the *complete* data. It is good to notice that this procedure has found the content vectors b^c also for the new utterance which is included in the *complete* data.

The model adaptation to the incomplete new style t (the target voice) can be done in closed form using the content vectors b^c learned during training. Suppose the aligned training data from our target speaker (style t) consists of M LSF vectors which by convention we considered to be in M different content classes $C = \{c_1, c_2, \dots, c_M\}$. We can derive the style matrix A^t that minimizes the total squared error over the target training data,

$$E^* = \sum_{c \in C} \|y^{tc} - A^t b^c\|^2. \quad (12)$$

The minimum of E^* is found by solving the linear system

$$\frac{\partial E^*}{\partial A^t} = 0. \quad (13)$$

The missing observations (LSFs) in the style t and a content class c of the test sentence can be then synthesized from $y^{tc} = A^t b^c$. This means we can estimate the target version of the test sentence by multiplying the target style matrix A^t with the content vectors corresponding to the test sentence. These vectors (b^c) were derived through the SVD training on the *complete* data of the source speakers.

2.4. Algorithm of the proposed method

The proposed technique is summarized in the following algorithm in which we assume that LSF features are available.

- 1) Time align the training data (source speakers and target speaker) and the test sentence (source speakers only) which is to be converted to the target voice. The alignment will respect the timeline or prosody of the main source speaker.
- 2) Form the *complete* data of the source speakers by combining their respective training data with their test sentence data.
- 3) Run the SVD to fit the asymmetric bilinear model to the *complete* data. This step will find the style matrices A^s for all the source speakers and the content vectors b^c for all the content (LSF) classes, including for those classes (LSFs) in the test sentence.
- 4) Find the style matrix A^t of the target voice by minimizing equation (12), thus by solving equation (13).

Synthesize the converted LSF vectors as $y^{tc} = A^t b^c$ with A^t found at step 4 and the content vectors b^c of the test sentence found at step 3.

3. Experiments

The work presented in this paper is concerned with spectral voice conversion. It focuses on the cases with limited training data, to demonstrate the beneficial properties that the proposed method has over the traditional GMM approach for voice conversion. Both objective measurements and listening tests are used in the evaluation.

In our experimental example there are two source speakers and one target speaker. We have used 16 kHz speech parameterized by the source plus filter speech representation proposed in [9] which models the spectral envelopes by 16-dimensional LSF vectors. As training data we have used a set of 3 sentences with maximized phonetic coverage uttered once by each of the three speakers (approx. 3s per utterance). The proposed method would work with only a single training sentence but it is known that the estimation of GMM parameters becomes unreliable on small datasets due to insufficient data. (Our previous paper [10] addressing the same issue of small datasets indicated that 3 utterances is the lower limit for a reliable estimation of GMM parameters and solved the problem of the reduced data sets by copying the covariance matrices of an existing GMM model.)

The method proposed in section 2.4 is compared against a modified GMM conversion method using 10 distinct test sentences. The GMM method was modified in a straightforward manner to take into account the presence of the second source speaker. The modified GMM model is estimating a joint pdf of the acoustic space of the 3 speakers (the 2 source speakers plus the target) based on the aligned training data.

The objective measurements involved distance measures in the feature domain between the converted LSFs and the real target LSFs. This required the prior alignment of the target test utterance and its converted (source) counterparts. We used DTW to align the source utterances to the target one respecting therefore the speaking rate of the target. This arrangement ensures that we obtain converted LSFs directly aligned to the target LSFs. This alignment to the target speaking rate served another purpose as the converted LSF could be readily replaced in the parameterization of the target test utterance and used with the remaining target parameters to synthesize a converted waveform. The converted waveforms therefore were synthesized from converted LSF parameters but left the other target parameters unchanged. These waveforms were evaluated in the listening tests mimicking the case where all other features were ideally converted to more effectively focus the evaluation on the performance of the spectral LSF conversion.

The objective results are summarized in Table 2. *Mean squared error* (MSE) was measured between the converted LSF vectors and the corresponding target vectors. The line spectral frequencies in each LSF vector range between 0 and 8000 Hz. Finally, the framewise MSE figures,

$$MSE(lsf_c, lsf_t) = \frac{\sum_{i=1}^{16} (lsf_c(i) - lsf_t(i))^2}{16}, \quad (14)$$

were averaged over the whole data. In the above equation, lsf_c and lsf_t denote the converted and target LSF vectors, respectively.

Spectral distortion (SD) was measured between a converted spectral envelope (derived from the converted LSF) and the corresponding target spectral envelope. The SD was measured

only for the perceptually relevant frequencies (only for the lower half of the spectrum), using

$$SD^2 = \frac{1}{(f_u - f_l)} \int_{f_l}^{f_u} \left(20 \log_{10} \left| \frac{H(e^{j2\pi f/f_s})}{\hat{H}(e^{j2\pi f/f_s})} \right| \right)^2 df, \quad (15)$$

where H and \hat{H} are the target and converted spectra, respectively, f_s is the sampling frequency, and f_l and f_u are the frequencies of integration (for better perceptual meaningfulness SD is computed between 0 and 4 kHz in our experiments).

Table 2: Results of the objective measurement tests showing MSE and SD scores

	Proposed Method	GMM
MSE	33303	38894
SD (dB)	5.36	5.65

Due to the possible sensitivity of the results to the choice of the training sets the objective measurements were repeated for a total of 4 different choices of the training set. The new measurements largely matched the figures already given in Table 2.

A listening test was also carried out using one fixed choice of the training set (the same training set as that used for generating the objective results shown in Table 2). The listening test was performed by 9 listeners. Each listener evaluated 10 sentences converted with both the GMM method and the proposed method. The listeners gave scores between 1 and 5 for each method and three different criteria: intelligibility, successfulness of identity conversion and overall sound quality. The results are given in Table 3 together with their 95% confidence intervals (with 1 indicating bad performance and 5 indicating excellent performance).

Table 3: Results of the subjective listening tests

	Proposed Method	GMM
Intelligibility	3.8±0.19	3.46±0.23
Identity	3.26±0.18	3.11±0.16
Quality	2.69±0.19	2.29±0.16

It is easy to see from Table 3 that the proposed method clearly outperforms the GMM based approach in this exemplary use case. The confidence intervals indicate a clear superiority of the proposed method in terms of overall speech quality. The advantage in identity conversion is not so evident but we can still conclude that the proposed method has at least the same performance in identity conversion as GMM.

4. Concluding remarks

In this paper we have presented a novel technique for voice conversion and studied its performance in the case of small training sets. Experimental results demonstrated its superiority to the GMM approach when dealing with the small datasets. The method however optimizes a similar mean squared error criterion function (equation 6) as the GMM conversion [2]. This explains why the proposed method suffers from the same over-smoothing problem as the GMM.

While this paper has presented just a simple implementation of the style-content framework for voice conversion, there is a wide potential for exploiting this framework further. Possible future directions for this work are to apply the style-content framework in cross-lingual voice conversion or to address the over-smoothing problem by using the proposed method on clustered acoustic space using the advantage that it copes better than GMM with reduced data. This study could also be

extended to the big training datasets where again a comparison with the GMM, or with some other voice conversion technique would be interesting.

5. Acknowledgements

This work was supported by the Academy of Finland, (application number 129657, Finnish Programme for Centres of Excellence in Research 2006-2011).

6. References

- [1] Tenenbaum, J. B. and Freeman, W. T., "Separating Style and Content with Bilinear Models", *Neural Computation* 12(6), 1247-1283, 2000
- [2] Kain, A. and Macon, M., "Spectral voice conversion for Text-to-Speech synthesis", in *Proceedings of International conference on Acoustics, Speech and Signal Processing*, Seattle, USA, 1998
- [3] Shuang, Z., Bakis, R. and Qin, Y., "Voice conversion based on mapping formants", in *Proceeding of TC-STAR Workshop on Speech-to-Speech Translation*, 2006, Barcelona, Spain
- [4] Narendranath, M., Murthy, H., Rajendran, S. and Yegnanarayana, N., "Transformation of formants for voice conversion using artificial neural networks", *Speech Communication*, vol.16, pp. 207-216, 1995
- [5] Kim, E. K., Lee, S. and Oh, Y., "Hidden Markov Model Based Voice Conversion Using Dynamic Characteristics of Speaker", In *5th Proceedings of European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997
- [6] Stylianou, Y., Cappe, O. and Moulines, E., "Continuous probabilistic transform for voice conversion", *IEEE Transaction on Speech and Audio Processing*, vol. 6, no.2, pp.131-142, 1998
- [7] Arslan, L. and Talkin, D., "Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum", In *5th Proceedings of European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997
- [8] 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
- [9] Nurminen, J., Popa, V., Tian, J. and Kiss, I., "A parametric approach for voice conversion", in *Proc. TC-STAR workshop on Speech-to-Speech Translation*, pp. 225-229, 2006
- [10] Jilei, T., Popa, V., and Nurminen, J., "Efficient Model Re-Estimation in Voice Conversion", *EUSIPCO 2008*, August 25 – August 29, 2008, Lausanne, Switzerland