A Parametric Approach for Voice Conversion

Jani Nurminen, Victor Popa, Jilei Tian, Yuezhong Tang, and Imre Kiss

Multimedia Technologies laboratory / Nokia Research Center
P.O. Box 100, FIN-33721 Tampere, FINLAND
{jani.k.nurminen, ext-victor.popa, jilei.tian, yuezhong.tang, imre.kiss}@nokia.com

Abstract

In voice conversion, speech and signal processing techniques are used for the modification of speaker identity, i.e. for modifying the speech of a source speaker to sound as if it was spoken by a target speaker. In this paper, we describe a parametric framework for voice conversion. The parametric representation separates the speech signal into a vocal tract contribution estimated using linear prediction and into an excitation signal modeled using a scheme based on sinusoidal modeling. This parametric framework is in line with the theory of human speech production and it also lends itself into very efficient compression. An initial version of the proposed voice conversion scheme has been implemented and evaluated in listening tests. The results show that the proposed approach offers a promising framework for voice conversion but further development work is still needed to reach its full potential.

1. Introduction

The term voice conversion refers to the modification of speaker identity by modifying the speech signal uttered by a source speaker to sound as if it was spoken by a target speaker. In general, a voice conversion system is first trained using speech data from both the source and the target speakers, and then the trained models can be used for performing the actual conversion. Potential applications for voice conversion include security related usage (hiding the identity of the speaker), entertainment applications, and text-to-speech (TTS) synthesis in which voice conversion techniques can be used for creating new and personalized voices in a cost-efficient way.

The research on voice conversion has received an increasing amount of attention, and a large number of different voice conversion approaches have been proposed in the literature. Roughly speaking, one way to categorize the voice conversion techniques is to consider three different aspects: the requirements concerning the training material (and the processing of this material), the domain of the conversion, and the techniques used for the actual conversion. For the training material, the most common approach is to require parallel speech from the speakers and to further align the data in a pre-processing step. Some papers, such as (Sundermann et al., 2004), have proposed techniques that relax this requirement by allowing the usage of different speech content from the two speakers. The ultimate goal in this type of work is to establish a good solution for cross-lingual voice conversion, i.e. for cases in which even the phoneme sets can be different for the source and the target speakers. Concerning the domain of conversion, many different methods have been reported in the literature, with the most common approach being a separate conversion of the vocal tract contribution and the excitation, as proposed e.g. in (Abe et al., 1988). There are also several different proposals for the modeling of the excitation signal in voice conversion. For example, in (Abe et al., 1988) the excitation was modeled in a very simple manner using only pitch and energy parameters, in (Kain & Macon, 1998) the source excitation was used with only pitch modification, while (Stylianou, Cappé & Moulines, 1998) used a more sophisticated harmonics + noise model. From the viewpoint of the actual conversion, example approaches presented in the literature include Gaussian mixture modeling (GMM) based conversion (Abe et al., 1998), neural network based conversion (Narendranath et al., 1995; Watanabe et al., 2002), hidden Markov model (HMM) based conversion (Kim et al., 1997), linear transformation based conversion (Ye & Young, 2003; Stylianou, Cappé & Moulines, 1998), and codebook based conversion (Arslan & Talkin, 1997).

The voice conversion system presented in this paper utilizes a parametric speech representation and operates on parallel training materials from the source and the target speakers. The conversion of the parameters is performed using a GMM based approach. Like the vast majority of the current voice conversion techniques, the conversion system presented in this paper focuses on the spectral aspects of conversion, including instantaneous pitch, and it does not modify the duration and timing related prosodic features or intonation contours. The most unique feature of the proposed voice conversion scheme is that it also enables very efficient speech coding seamlessly within a single framework.

The rest of this paper is organized as follows. First, the parametric speech model used in the voice conversion system is presented in Section 2. Then, in Section 3, we describe the voice conversion scheme, including aspects related to both the training phase and the actual conversion. Section 4 discusses the results obtained in the second TC-STAR evaluation campaign (see the TC-STAR web page, www.tc-star.org for more detailed information). Finally, some concluding remarks are presented in Section 5.

2. Parametric speech model

The voice conversion scheme presented in this paper is based on a parametric speech model. The selection of such a model was inspired by the successful usage of a similar model in a low-bit-rate speech coding application, presented earlier in (Rämö et al., 2004). The parametric model described in this section contains favorable properties from the viewpoint of both voice conversion and speech coding, and allows a seamless combination of these two aspects.

2.1. Speech representation

The speech model used in this paper is based on the fact that a speech signal, or alternatively a vocal tract

excitation signal, can be represented as a sum of sine waves of arbitrary amplitudes, frequencies and phases (McAulay & Quatieri, 1986; McAulay & Quatieri, 1995):

$$s(t) = \mathrm{Re}\sum_{m=1}^{L(t)} a_m(t)\exp(j\left[\int_0^t \omega_m(t)dt + \theta_m\right]), \quad (1)$$

where, for the mth sinusoidal component, $a_m(t)$ and $\omega_m(t)$ represent the amplitude and the frequency, and θ_m represents a phase offset. To obtain a frame-wise representation, the parameters are assumed to be constant over the analysis frame. Consequently, the discrete signal $s(n)$ in a given frame can be approximated as

$$s(n) = \sum_{m=1}^{L} A_m \cos(n\omega_m + \theta_m), \quad (2)$$

where A_m and θ_m represent the amplitude and the phase of each sine-wave component associated with the frequency track ω_m, and L denotes the number of sine-wave components.

To simplify the representation, we have assumed that the sinusoids are always harmonically related, i.e. that the frequencies of the sinusoids are integer multiples of the fundamental frequency ω_0. During voiced speech, ω_0 corresponds directly to the pitch associated with the analysis frame. During unvoiced speech, however, there is no physically meaningful pitch available, and we use a fixed value for ω_0. To further simplify the model, we assume that the sinusoids can be classified as continuous or random-phase sinusoids. The continuous sinusoids represent voiced speech and they are modeled using a linearly evolving phase. The random-phase sinusoids, on the other hand, represent unvoiced noise-like speech that is modeled using a random phase.

To facilitate both voice conversion and speech coding, the sinusoidal model described above is applied to the modeling of the vocal tract excitation signal. The excitation signal is obtained using the well-known linear prediction approach. In other words, the vocal tract contribution is captured by the linear prediction analysis filter $A(z)$ and the synthesis filter $1/A(z)$, while the excitation signal is obtained by filtering the input signal $x(t)$ using the linear prediction analysis filter $A(z)$ as

$$s(t) = x(t) - \sum_{j=1}^{N} a_j x(t-j), \quad (3)$$

where N denotes the order of the linear prediction filter. In addition to the separation into the vocal tract model and the excitation model, the overall gain or energy is used as a separate parameter to simplify the processing of the spectral information.

2.2. Parameter estimation

As described in Section 2.1, the speech representation used in the voice conversion system consists of three elements: i) vocal tract contribution modeled using linear prediction, ii) overall gain/energy, iii) normalized excitation spectrum. The third of these elements, i.e. the residual spectrum, is further represented using the pitch, the amplitudes of the sinusoids, and voicing information. Each of these parameters is estimated at 10-ms intervals from 8-kHz input speech signal. The rest of this section describes the parameter estimation process at a general level.

The coefficients of the linear prediction filter are estimated using the autocorrelation method and the well-known Levinson-Durbin algorithm, together with mild bandwidth expansion. This approach ensures that the resulting filters are always stable. Each analysis frame consists of a 25-ms speech segment, windowed using a Hamming window. The degree of the linear prediction filter is set to 10 for 8-kHz speech. For further processing, the linear prediction coefficients are converted into the line spectral frequency (LSF) representation. From the viewpoint of voice conversion, this widely-used representation is very convenient since it has a close relation to formant locations and bandwidths, and it offers favorable properties for different types of processing and guarantees filter stability.

The operation of the pitch estimation algorithm can be summarized as follows. First, a frequency-domain metric is computed using a sinusoidal speech model matching approach that partially follows the ideas presented in (McAulay & Quatieri, 1990). Then, a time-domain metric measuring the similarity between successive pitch cycles is computed for a fixed number of pitch candidates that received the best frequency-domain scores. The actual pitch estimate is obtained using the two metrics together with a pitch tracking algorithm that considers a fixed number of potential pitch candidates for each analysis frame. As a final step, the obtained pitch estimate is further refined using a sinusoidal speech model matching based technique to achieve better than one-sample accuracy.

Once the final refined pitch value has been estimated, the parameters related to the residual spectrum can be extracted. For these parameters, the estimation is performed in the frequency domain after applying variable-length windowing and fast Fourier transform (FFT). The voicing information is first derived for the residual spectrum through the analysis of voicing-specific spectral properties separately at each harmonic frequency. The spectral harmonic amplitude values are then computed from the FFT spectrum. Each FFT bin is associated with the harmonic frequency closest to it.

The final step in the parameter estimation process is to obtain the energy value. This estimation is performed in time domain, using the root mean square energy. Since the frame-wise energy varies significantly depending on how many pitch peaks are located inside the frame, the estimation computes the energy of a pitch-cycle length signal instead.

3. Parametric conversion scheme

Our voice conversion system performs the conversion using the parametric representation presented in Section 2. The GMM based conversion models are trained using parallel aligned data from the source and target speakers. The first part of this section describes the techniques used for the data alignment and model training, while the latter part describes the actual models and their usage in the conversion phase.

3.1. Alignment and training

The training of the GMM based model utilizes aligned parametric data from the source and target voices. The

alignment is achieved in two steps. First, both the source and target speech signals are segmented and then a finer-level alignment is performed within each segment. The segmentation is performed at phoneme-level using HMM models, and the alignment utilizes dynamic time warping (DTW). It is also possible to utilize manually labeled phoneme boundaries if such information is available but this is not used as the only solution to avoid the requirement for any manual processing that would anyway be time-consuming and prone to human errors.

In principle, the speech segmentation could be conducted using very simple techniques, for example by measuring spectral change without taking into account knowledge about the underlying phoneme sequence. However, to achieve better performance, we fully exploit the information about the phonetic content and perform the segmentation using HMM based models. The first step is to extract a sequence of feature vectors from the speech signal. The extraction is performed frame by frame, using similar frames as in the parameter extraction procedure described in Section 2. The phoneme sequence associated with the corresponding speech is assumed known. Given the phoneme sequence, a compound HMM model is built up by sequentially concatenating the phoneme HMM models. Next, the frame-based feature vectors are associated with the states of the compound HMM model using Viterbi search to find the best path (Rabiner & Juang 1993). By keeping track of the states, a backtracking procedure can be used to decode the maximum likelihood state sequence. The phoneme boundaries in time are then recovered by following the transition change from one phoneme HMM to another.

The phoneme-level alignment obtained using the procedure above is further refined by performing frame-level alignment using dynamic time warping. DTW is a dynamic programming technique that can be used for finding the best alignment between two acoustic patterns. This is functionally equivalent to finding the best path in a grid to map the acoustic features of one pattern to those of the other pattern. Finding the best path requires solving a minimization problem, minimizing the dissimilarity between the two speech patterns. In our implementation, DTW is applied on Bark-scaled LSF vectors and the algorithm is constrained to operate within one phoneme segment at a time. Non-simultaneous silent segments are disregarded.

The DTW algorithm results in a combination of aligned source and target vectors $\mathbf{z}=[\mathbf{x}^T\ \mathbf{y}^T]^T$ that can be used to train a conversion model. In the training, we have used the popular approach proposed in (Kain & Macon, 1998) that makes use of the aligned data $\mathbf{z}$ to estimate the GMM parameters (α, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) of the joint distribution p($\mathbf{x}$,$\mathbf{y}$). This is accomplished iteratively through the well-known Expectation Maximization (EM) algorithm (Dempster, Laird & Rubin 1977). In the above notation, $\mathbf{x}$ and $\mathbf{y}$ correspond to the source and target feature vectors, respectively.

3.2. Conversion

Among the speech parameters described in Section 2, pitch and LSFs were found particularly important from the perception point of view in voice conversion. In the development of our current system, the emphasis was placed on the conversion of these features. Other features such as voicing and residual spectrum were used as complementary information and were exploited in the model training but no explicit conversion was performed for these parameters in the current system.

The conversion of the speech parameters follows a scheme where the trained GMM parameterizes a linear function that minimizes the mean squared error (MSE) between the converted source and target vectors. The conversion function is, as shown in (Kain & Macon, 1998):

$$F(\mathbf{x}) = E(\mathbf{y} \mid \mathbf{x}) = \sum_{i=1}^{L} p_i(\mathbf{x}) \cdot \left(\boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx} \left(\boldsymbol{\Sigma}_i^{xx} \right)^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_i^x \right) \right), \quad (4)$$

where

$$p_i(\mathbf{x}) = \frac{\alpha_i \cdot N(\mathbf{x}, \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx})}{\sum_{j=1}^{L} \alpha_j \cdot N(\mathbf{x}, \boldsymbol{\mu}_j^x, \boldsymbol{\Sigma}_j^{xx})}. \quad (5)$$

The covariance matrix is formed as

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy} \\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy} \end{bmatrix}, \quad (6)$$

and

$$\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_i^x \\ \boldsymbol{\mu}_i^y \end{bmatrix}, \quad (7)$$

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

The conversion of the LSF vectors is performed using an extended vector that also contains the derivative of the LSF vector, to take some dynamic context information into account. This combined feature vector is transformed through GMM modeling, using Equation (4). Only the true LSF part is retained after conversion. The conversion utilizes several modes, each containing its own GMM model with 8 mixtures. The modes are achieved by clustering the LSF data in a data-driven manner.

The pitch parameter is transformed through the associated GMM in frequency domain using Equation (4). During unvoiced parts, "pitch" is left unchanged. The 8-mixture GMM used for pitch conversion is trained on aligned data, with a requirement to have matched voicing between the source and the target data.

After the conversion of the pitch parameter, the residual amplitude spectrum is processed accordingly. The reason for this processing is the fact that the length of the amplitude spectrum vector depends on the pitch value at the corresponding time instant. This means that the residual spectrum, although essentially unchanged, will be re-sampled to fit the dimension dictated by the converted pitch at that time.

After the features have been converted, they are used together to re-synthesize the transformed waveform. The synthesis is performed in a pitch-synchronous manner.

4. Evaluation results

The parametric voice conversion system described in this paper was evaluated in listening tests in the context of the second TC-STAR evaluation campaign. The evaluation

covered aspects related to both speaker identity and speech quality. The evaluation was carried out by an independent evaluation agency.

4.1. Test set-up

The data set used in the testing included UK English speech data from four different speakers (two female and two male speakers). The training set included 159 sentences per speaker and a distinct testing set consisted of 9 sentences per speaker. The same sentences were recorded from all the speakers.

Among the 12 possible conversion directions, 4 were chosen as the directions included in the test. For the selected directions, the test organizer provided the recorded source sentences used in the test. These source sentences were converted using our voice conversion system to the voices of the target speakers. The converted signals were evaluated by 20 native non-expert listeners.

The listening test included two parts. In the first part, the listeners were asked to evaluate the speaker identity without considering the speech quality using the 5-level scale summarized in Table 1. The true target signals recorded from the target speakers, available only for the test organizer, were used as the reference. In the second part, the listeners evaluated the perceptual quality of the converted speech using the mean opinion score (MOS) grades shown in Table 2.

Table 1. Scale used for evaluation of speaker identity. The listeners were asked to evaluate whether the two samples in the given pair were spoken by the same person or not.

Grade	Meaning
5	Definitely identical
4	Probably identical
3	Not sure
2	Probably different
1	Definitely different

Table 2. Scale used in the evaluation of speech quality.

Grade	Meaning
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

4.2. Results

The results are summarized in Table 3 and Table 4. Table 3 contains the results from the first part of the listening test, focusing on the evaluation of speaker identity. The combined score for all the directions, not shown in the table, was 2.53. The results from the speech quality evaluation are summarized in Table 4.

Table 3. Results from the first part of the evaluation (speaker identity). F denotes a female and M a male speaker.

Direction	$F_1 \rightarrow F_2$	$F_1 \rightarrow M_2$	$M_1 \rightarrow F_2$	$M_1 \rightarrow M_2$
Score	3.10	3.05	2.20	1.77

Table 4. Results achieved from the second part of the evaluation (speech quality).

	MOS score
Achieved score	2.09
Reference 1 (source)	4.80
Reference 2 (target)	4.78

4.3. Discussion on the results

When looking at the results, the first observation that can be made is that there were large differences between the different conversion directions. Moreover, despite the moderate average scores, the person identity conversion was sometimes perceived very successful. This can be regarded as a good result due to two reasons. First, our initial system that participated in the evaluation only converted the LSFs and the pitch parameter. Moreover, the conversion was performed in a frame-wise manner without considering the frame-to-frame evolvement of the parameters or intonation contours. Significant improvements can be expected after making the system more complete.

As can be seen from Table 4, a rather low score was achieved in the speech quality evaluation. There are a couple of clear reasons for this. First, the system produced 8-kHz output signals while the other signals (e.g. the reference samples) included in the listening test used a sampling rate of 16 kHz. Second, the source signals also contained some non-speech elements such as audible breathing and the parametric speech and conversion models created many audible artifacts to the corresponding places in the output signals. Third, the frame-by-frame conversion made the converted parameter contours a bit noisy and this was also audible in the output signals. Finally, the fact that not all the parameters were converted also had its impact on the quality. Considering these underlying reasons for quality degradations, it is evident that much better quality can be expected after further development work.

5. Concluding remarks

This paper has introduced a parametric approach for voice conversion. The parametric representation both facilitates the conversion and enables very efficient speech compression. The proposed model represents the vocal tract contribution using line spectral frequencies and the excitation signal is modeled using a scheme based on sinusoidal modeling. The actual conversion is performed using a GMM based approach. The practical evaluation results support our assumption that the proposed approach is very promising but more development work is still needed to achieve its full potential.

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

Abe, M., Nakamura, S., Shikano, K., Kuwabara, H. (1988). Voice conversion through vector quantization. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'88*, New York, NY, USA, pp. 655-658.

Arslan, L.M., Talkin, D. (1997). Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum. In *Proceedings of Eurospeech'97*, Rhodes, Greece, pp. 1347-1350.

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

Kain, A., Macon, M.W. (1998). Spectral voice conversion for text-to-speech synthesis. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'98*, Seattle, WA, USA, pp.285-288.

Kim, E.K., Lee, S., Oh, Y.-H. (1997). Hidden Markov Model Based Voice Conversion Using Dynamic Characteristics of Speaker. In *Proceedings of Eurospeech'97*, Rhodes, Greece, pp. 2519-2522.

McAulay, R.J., Quatieri, T.F. (1986). Speech Analysis/Synthesis based on a Sinusoidal Representation. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-34, No. 4, pp. 744-754.

McAulay, R.J., Quatieri, T.F. (1990). Pitch Estimation and Voicing Detection based on a Sinusoidal Speech Model. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP'90*. pp. 249-252.

McAulay, R.J., Quatieri, T.F. (1995). Sinusoidal Coding. In W.B. Kleijn, K.K. Paliwal (Eds.), *Speech Coding and Synthesis*. Elsevier Science B.V., pp. 121-174.

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

Rabiner, L. and Juang, B.-H. (1993). Fundamentals of speech recognition, Prentice-Hall, USA.

Rämö, A., Nurminen, J., Himanen, S., Heikkinen, A. (2004). Segmental Speech Coding Model for Storage Applications. In *Proceedings of Interspeech 2004 ICSLP International Conference on Spoken Language Processing*. Jeju Island, South Korea.

Stylianou, Y., Cappe, O., Moulines, E. (1998). Continuous probabilistic transform for voice conversion. IEEE Transactions on Speech and Audio Processing, Vol. 6, No.2, pp.131-142.

Sundermann, D., Bonafonte, A., Ney, H., Hoege H. (2004). A first step towards text-independent voice conversion. In *Proceedings of Interspeech 2004 ICSLP International Conference on Spoken Language Processing*, Jeju Island, South Korea.

Watanabe, T., Murakami, T., Namba, M., Hoya, T., Ishida, Y. (2002). Transformation of spectral envelope for voice conversion based on radial basis function networks. In *Proceedings of Interspeech 2002 ICSLP International Conference on Spoken Language Processing*, Denver, USA, pp. 285-288.

Ye, H., Young, S. (2003). Perceptually weighted linear transformations for voice conversion. In *Proceedings of Eurospeech'03*, Geneva, Switzerland.