Parametric Speech Coding Framework for Voice Conversion Based on Mixed Excitation Model

Michał Lenarczyk

Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warsaw, Poland

Abstract. Adaptation of mixed-excitation linear predictive (MELP) model for application in voice conversion is presented. The adapted model features only numerical parameters which can be used for phonetic space transformation from source to target speaker using methods of machine learning. The validity of the model was demonstrated by applying transformation to both the pitch and the spectral envelope of voice.

1 Introduction

Voice conversion is a technique of transforming speaker individuality in speech signal. The aim is to change voice of an original (source) speaker to sound like another (target) speaker, while preserving the semantic content of the utterance. A voice conversion system follows the general structure of Fig. 1.

Signal processing techniques are applied in analysis and synthesis stages, to transform speech signal to and from a parametric representation that is suitable for transformation, which is most commonly done using machine learning techniques. The analysis and synthesis can thus be viewed as a pre- and post-processing for representing data in a concise form, such that transformation function can be estimated given speech data of the source and target speakers. One approach to learning the source-target transformation originally introduced in the classic work of Abe et al. [1] is to use parallel corpora (composed of the same set of utterances from both speakers) and extract numerical features from time aligned source and target speech signals. In this case, there are a number of general purpose methods that can be applied for transformation. They include, among others: frequency warping functions [2], maximum likelihood modeling including gaussian mixture models (GMM) [3,4], hidden Markov models (HMM) [5,6], support vector regression [9], or artificial neural networks (ANN) [8,7]. On the other hand, the representation of speech (coding) is typically defined differently by different authors, making it difficult to evaluate fitness and compare particular transformation methods in the task of interest.

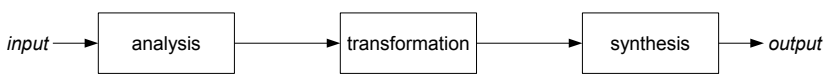

Fig. 1. General structure of a voice conversion system

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 507–514, 2014.

Considering the chain of Fig. 1 with the transformation function removed, i.e. replaced by identity function (channel), we recognize it as a prototype vocoder. Rather than building a VC system from scratch, use of one of established voice coding methods is therefore possible. In this paper, adaptation of the mixed excitation linear predictive (MELP) coding is presented that lends itself for use in selective transformation of voice features.

There are a number of speech features that define speaker individuality. The instantaneous configuration of formants in signal defines the phoneme uttered but the overall characteristics of the frequency envelope of speech over entire acoustic space is the individual voice, or *timbre*. Pitch, or perceived fundamental frequency in a signal, is the primary distinguishing feature of male and female voices. Accentuation and prosodic features such as the modulation of pitch and amplitude in time, duration of phonemes, etc. also count as characteristic properties of speakers. It is generally agreed that timbre and pitch are the most important and they are the focus of this work.

2 Parametric Modeling of Speech

The mixed excitation linear predictive (MELP) coder, originally introduced in [10], is an example of a highly (yet not fully) parametric coding scheme. The basis of this coder is the linear predictive coding (LPC), which models the speech production as a source-filter process. It captures the general shape of speech spectrum (envelope), which corresponds roughly to the effect of vocal tract and lips, in the form of a filter defined by a small number of coefficients. The model residual is believed to represent phonation, be it voiced (harmonic) or unvoiced (noisy). MELP models the excitation as a combination of harmonic pulse train and white noise, mixed in appropriate proportion.

The initial motivation for MELP was low bitrate coding. As such, it features quantization and several other nonparametric aspects that are an obstacle for manipulation, such as switched voicing state, switched pitch aperiodicity, noise thresholds, and unfavourable spectral envelope representation. This section briefly describes the structure of the vocoder and details changes introduced into the adapted model. As a baseline, the 2.4 kb/s MELP specification accepted as the U.S. Federal Standard is taken [12].

Modification of parameters can be done with any method. Preliminary experiments conducted with static and neural network-learned transformations are described in Section 3.

2.1 Structure of MELP

The foundation of MELP synthesis is a mixed source composed of white noise and a harmonic component represented by a sequence of pulses at defined pitch intervals, followed by an all-pole synthesis filter. The two basic types of excitation are combined in a harmonic/noise proportion that corresponds to one of three allowed states of voicing: fully voiced (0.8/0.2), weakly voiced (0.5/0.5) and unvoiced (0.0/1.0), with interpolation between the two voiced states. On top of this basic structure, a number of additional features have been added to enhance speech quality while keeping the bit rate low. They include:

1. switched aperiodicity that adds jitter to the harmonic excitation, applied in the weakly voiced case,
2. encoded amplitudes of initial harmonics to better represent the characteristics of the periodic excitation,
3. two conditioning filters for shaping pulse and noise spectra (boost low frequency part of harmonic spectrum and high frequency part of noise spectrum to approximate the behaviour observed in speech) while assuring spectral flatness of the excitation mixture,
4. subband voicing analysis allowing finer shaping of the excitation's harmonic to noise ratio with a staircase approximation,
5. perceptual weighting filter in the postprocessing phase for formant sharpening (adaptive spectral enhancement).

2.2 Proposed Vocoder Framework

The vocoder elaborated by the author is proposed as a flexible and fully parametric framework for application in voice conversion that builds on the mixed-excitation model. The structure preserves essential parts of the original MELP coder and modifications were designed to make the speech representation suitable for learning and transformation.

The framework has been adapted to process speech signal sampled with 16 kHz as opposed to 8 kHz used in telephone-band speech coding. This adds an additional octave that is important in fricatives. The frame size is set to 10 ms (160 samples) as opposed to 22.5 ms in MELP, and analysis is performed in 30 ms window (480 samples), as opposed to 25 ms.

The synthesis section (Fig. 2-b) preserves the mixed excitation composed of white noise and periodic pulse with period given by the pitch lag parameter. Envelopes for both source types are generated by dividing the overall gain into harmonic and noise parts according to voicing level parameter, and interpolating with a filter based on a Hamming-windowed sinc function truncated at ± 320. The mixer shapes the periodic and noise signals using the filters $H_P(z) = 1 + az^{-1}$ and $H_S(z) = 1 - bz^{-1}$, respectively, with b fixed at 1 and a calculated according to $a = \frac{G_S^2}{G_P^2}$ where G_P, G_S denote pulse and noise gain. This is equivalent to the formulations in [10] where a different gain scheme is used. The process of generating the excitation can be performed in subbands. The synthesis filter is lattice IIR of order 18. The parameters for synthesis are estimated in the analysis section (Fig. 2-a) which differs from baseline MELP both in representation and methods used. The following paragraphs detail the differences.

Preprocessing: a first order preemphasis is applied during preprocessing to compensate the spectral tilt. This preconditioning is sufficient to accurately model the entire spectrum and thus adaptive spectral enhancement is not employed.

Spectral envelope representation: spectral envelope is represented by filter estimated using standard autocorrelation method of LPC and the coefficients are found in the Levinson-Durbin procedure. The LP filter order is increased from 10 to 18 to account for the wider frequency band.

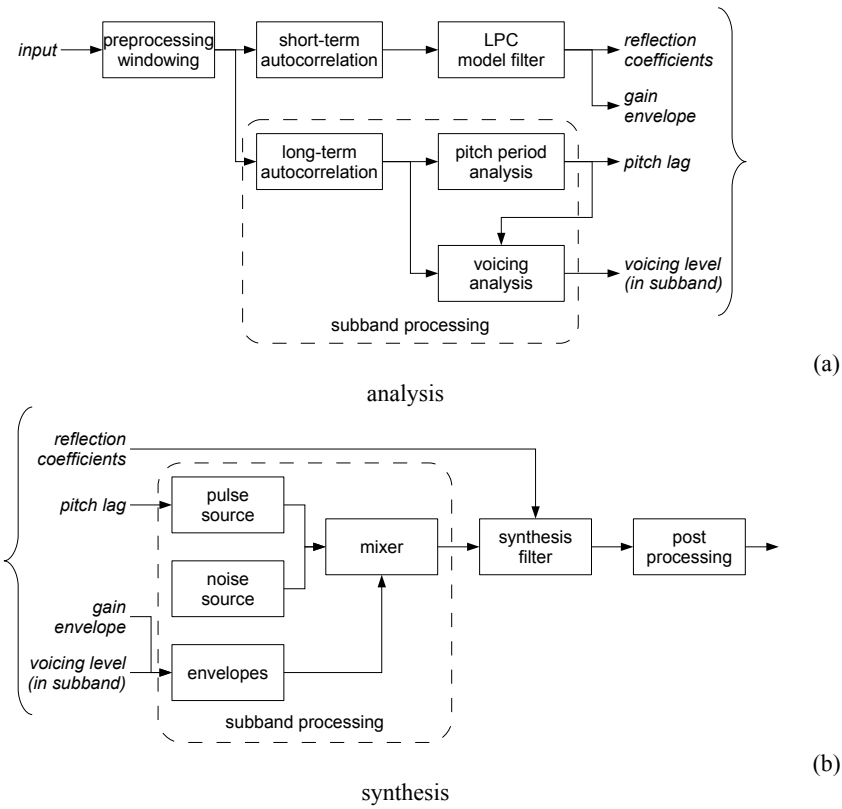

Fig. 2. Structure of the parametric framework.

Representation of the envelope is of primary importance. The proposed framework abandons line spectrum frequencies (LSF) representation, as used in most modern speech coders, in favor of reflection coefficients of a lattice filter. LSF have favourable quantization and interpolation properties, but require strict ordering to preserve filter stability which may be hard to assure when transforming using general purpose methods of machine learning. On the other hand, stability can be easily assured by bounding the reflection coefficients to the range $k_i \in (-1, +1)$. In experiments conducted with neural network based transformation, the requirement for restricted range was naturally met by selecting a bipolar activation function. If other methods of transformation are to be used, log area ratios (LAR) given by $\log \frac{1-k_i}{1+k_i} \in \mathbb{R}$ may be a preferable representation. Interpolation in reflection coefficient domain gives poor results and is thus avoided by reducing the frame length to 10 ms.

Correction of harmonic amplitudes: in the standard MELP [12], Fourier magnitudes of initial harmonics of the LPC residual of voiced speech were added because it is known [13] that LPC fails to model harmonic spectra accurately. Because harmonic frequencies change when pitch is altered, and the residual values are not predictable

from the model, they are a non-parametric feature and are thus omitted. Study on relative importance of parameters in bitrate-reduced version of MELP [11] revealed Fourier magnitudes to be of secondary importance for quality.

Pitch: the pitch period is evaluated by searching the position of the dominant peak of autocorrelation function in a defined interval. To increase reliability, the original and $2\times$ upsampled and interpolated autocorrelation vectors are upper halfwave rectified and multiplied to produce a periodicity index, which peaks at double the fundamental period. It was found to be a good estimator with a half sample resolution. Explicit checks are additionally performed to avoid selecting halved or doubled pitch period. Aperiodicity (pitch jitter) is not modeled.

Voicing estimation: the voicing level is estimated as a continuous fraction of the harmonic component in the overall mixture. This is in contrast with MELP where the level of voicing is nominal with three possible levels (fully voiced, jittery voiced and unvoiced). The associated ratios of harmonic to total energy are fixed at 0.8, 0.5 and 0, respectively, with interpolation between the voiced states. Such a coding produces a degenerate distribution which is hard to model using machine learning. The proposed method avoids quantization of voicing which reduces the amount of logic and allows gradual transformation towards breathy or whispered speech.

The evaluation of this proportion is based on a weighted sum of peak levels of autocorrelation at the pitch lag and its multiples. The value is always clipped to the range $[0, +1]$. Typically, it is close to 1 for fully voiced and slightly above 0 for unvoiced speech, i.e. there remains some residual energy of the other kind. Thich is not a problem : in voiced phonation, the additional noise adds naturalness and in unvoiced speech the weak harmonic component is dominated by noise. Moreover, in unvoiced segments the pitch estimate tends to fluctuate, making the impulse train less regular and easier to vanish in noise. Fig. 3 illustrates the method with a sample speech utterance.

Gain envelopes: the gain is estimated every 10 ms from residual energy over the window. In the synthesis stage, the vocing level is used to divide the energy into harmonic/stochastic parts, and smooth envelopes are generated at the sample rate by interpolating in log domain using Hamming windowed sinc function filter.

2.3 Subband Analysis

For a more realistic modeling of phonemes that are both noisy and harmonic, but with a different ratio depending on frequency region, subband processing was proposed in [10] and is followed in this work. For this mode of analysis, four frequency bands are defined and the level of voicing is estimated from band filtered speech. When generating excitation, each voicing level is used to generate a band limited mixture with a different proportion of harmonic and noisy components. The subband excitations are finally added to form a total excitation that preserves spectral flatness. As shown on Fig. 4, higher frequency bands tend to be more noisy when the overall voicing is high.

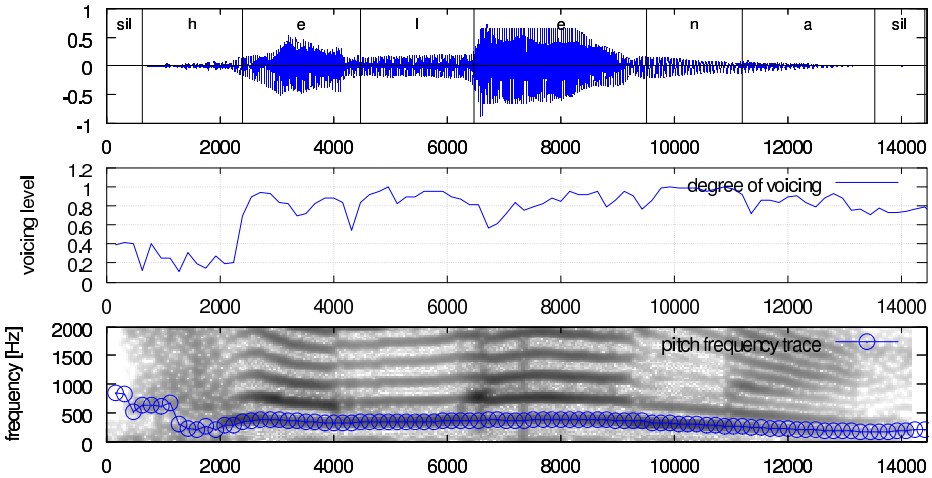

Fig. 3. Pitch tracking and voicing level detection for the word "Helena" uttered by a boy. From top to bottom: original speech waveform, voicing level estimate, spectrogram with pitch frequency estimate overlaid.

3 Transcoding Experiments

To evaluate the quality of the coding method and its applicability for conversion of voice characteristics, transcoding was applied to a small subset of samples from the *Corpora* Polish speech corpus [14]. Informal listening tests were conducted in three conditions: no transformation, pitch and voicing scaling and spectral envelope transformation. This

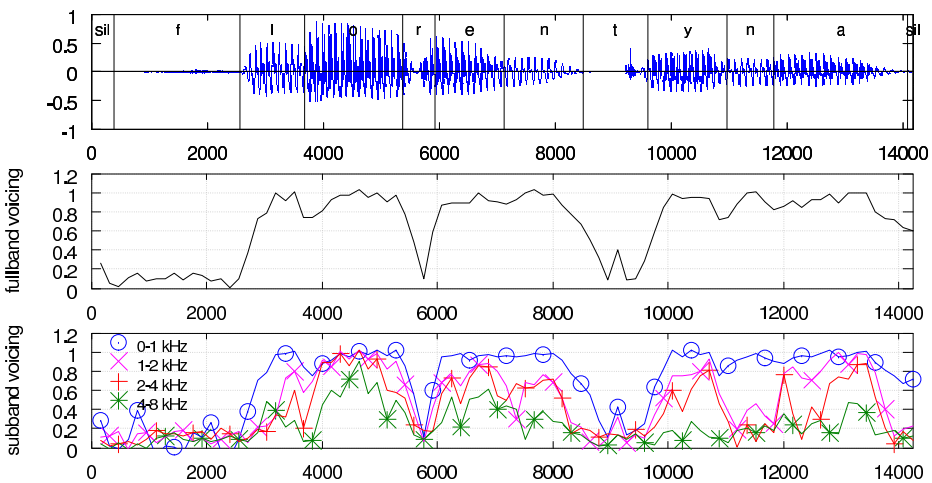

Fig. 4. Male utterance "Florentyna" (top) and associated voicing traces from full band (middle) and subband (bottom) analysis. Four subbands are used: 0–1 kHz; 1–2 kHz; 2–4 kHz; 4–8 kHz.

evaluation is a preliminary proof of validity of the framework and formal evaluation is planned. The obtained samples can be accessed from the author's website [15].

In the first scenario, the parameters were extracted and used directly for speech reconstruction. This provides a baseline estimate for the level of quality achievable through this coding scheme. Two versions were tested : with a single level of voicing evaluated over the full frequency band, and with voicing calculated over four subbands (0–1 kHz, 1–2 kHz, 2–4 kHz and 4–8 kHz). In both cases, different kind of speech degradation is audible. Fullband processing yields a buzzy quality of voice especially in male speech. Subband approach offers a voice quality that is very close to the source but not transparent, and suffers from occasional tonal thumps that are attibuted to less reliable estimation of voicing, which results in noise bursts (cf. Fig. 4). This effect is especially pronounced at plosives. Better subband pitch tracking may help reduce the problem.

The second scenario is concerned with pitch modification, which was done by rescaling the frequency estimate by a predefined ratio. Informal listening indicates that a considerable amount of scaling (by a factor of two, or one octave, up or down) can be applied without introducing additional distortion compared to unscaled pitch. Similar modification of voicing was also performed and demonstrates the ability to generate breathy and whispered speech.

In the last scenario, the frequency envelope of speech was transformed toward a desired target. To achieve this, the reflection coefficients of the model filter were transformed with the use of a neural network. The network had a layered structure with one hidden layer of 40 units and input and output layers each of 18 units which is the size of the reflection coefficients vector. A bipolar activation function was selected and the network was trained using backpropagation algorithm with a learning constant of 0.0001. A parallel set of over 10000 input-output vectors was created from 251 short utterances (names, numbers, control words) of the speech corpus. In repeated experiments, the learning process consistently led to a 10 to 20-fold reduction in error within initial 100 iterations, around 10% reduction within the next 100 iterations, and no appreciable further gain. The error contribution per vector stabilised at around 0.5.

As a result of transformation of the vocal tract parameters, the speech underwent definite qualitative transformation. The most obvious artifact in the resynthesised samples is a muffling of voice. This muffling is presumably due to variability in pronounciation of phonemes across utterances and the fact that alignment was only based on phonetic annotation and not on numerical procedures like the commonly used dynamic time warping (DTW). Another possible cause is the attraction of filter polynomial roots toward the center of the Z-plane ($z = 0$) as a result of averaging, which causes loss of formant sharpness.

4 Concluding Remarks

A parametric speech coding framework has been proposed allowing selective transformations of voice properties. Modification of pitch, voicing and frequency envelope was demonstrated. With direct transcoding the obtained quality was good but not transparent, which puts a limit on quality with conversion applied. Solution to some issues (e.g.

subband mode thumps) seems possible. On the other hand, the framework has the potential for the emerging field of real time voice conversion. Further work is directed towards this goal and formal listening tests are planned.

Acknowledgement Study was supported by research fellowship within "Information technologies: research and their interdisciplinary applications" agreement number POKL.04.01.01-00-051/10-00.

References

1. Abe M., Nakamura S., Shikano K., Kuwabara H., Voice conversion through vector quantization, *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1 pp. 655–658, 1988
2. Hanzlíček Z., Matoušek J., On using warping function for LSFs transformation in a voice conversion system, *Proc. Int. Conf. Signal Processing*, pp. 2725–2728, 2008
3. Stylianou, Y., Cappé, O., Moulines, E., Continuous probabilistic transform for voice conversion, *IEEE Transactions on Speech and Audio Processing*, vol. 6 no 2, pp. 131–142, 1998
4. Ye, H., Young, S., Quality-enhanced voice morphing using maximum likelihood transformations, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14 no 4, pp. 1301–1312, 2006
5. Arslan, L. M., Talkin, D., Speaker transformation using sentence HMM based alignments and detailed prosody modification, *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, pp. 289–292, 1998
6. Rentzos, D., Vaseghi, S., Yan Q., C. Ho, Parametric Formant Modelling and Transformation in Voice Conversion, *International Journal of Speech Technology*, vol. 8 no 3, pp. 227–245, 2005
7. Guido, R. C., Sasso Vieira, L. Barbon J., S. Sanchez, F. L. Maciel, C. D., Everthon S. F., Pereira, C. J., A Neural-wavelet Architecture for Voice Conversion *Neurocomputing*, vol. 71 no 1-3, pp. 174–180, 2007
8. Orphanidou, C., Moroz, I. M., Roberts, S. J., Wavelet-based voice morphing, *WSEAS Journal on Systems*, vol. 10, pp. 3297–3302, 2004
9. Laskar R., Talukdar F., Bhattacharjee R., Das S., Voice Conversion by Mapping the Spectral and Prosodic Features Using Support Vector Machine, *Applications of Soft Computing*, vol. 28, 2009
10. McCree A., Barnwell, T. P., A new mixed excitation LPC vocoder, *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, pp. 593–596, 1991
11. McCree, A., De Martin, J. C., A 1.6 kb/s MELP coder for wireless communications, *Proc. IEEE Workshop on Speech Coding for Telecommunications*, pp. 23–24, 1997
12. Supplee L. M., Cohn R. P., Collura J. S., McCree A. V., MELP: the new Federal Standard at 2400 bps, *Proc. Int. Conf. Acoustics, Speech and Signal Processing* vol. 2 pp. 1591–1594, 1997
13. Makhoul J., Linear prediction: A tutorial review, *Proc. IEEE*, vol. 63 no 4 pp. 561–580, 1975
14. Grocholewski, S. CORPORA - speech database for Polish diphones. *Proc. EUROSPEECH*, 1997
15. Author's web page: `http://phd.ipipan.waw.pl/~m.lenarczyk/TSD2014.html`