Voice Conversion and the Prospects of Real Time Applications

Michał Lenarczyk

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

Abstract. A voice conversion system is presented that is based on an original parametric vocoder that allows real time speech transcoding with a minimal latency. The vocoder is based on the mixed excitation paradigm. Instead of the error-prone voiced-unvoiced frame classification, a continuous degree of voicing is employed, which is interpreted as the ratio of harmonic to total signal energy. Subband-based estimation is used for improved modeling of the degree of voicing across frequency. A training set of parameters obtained from a parallel speech corpus was used to obtain a conversion function by training neural network regression model. Processed speech signals were evaluated in formal listening tests. Results indicate a successful conversion but also a significant loss in quality.

1 Introduction

This article concerns real time implementation of voice conversion. Voice conversion is a speech processing technique which changes the perceived identity of voice to that of a target speaker, while preserving the structural content of the utterance. The aim is to not only conceal the personality of the original speaker, but to closely adapt characteristics of speech to the target voice, to make a convincing impression that it was uttered by a designated speaker.

Speech signal can generally be considered to contain three types of information : linguistic information, which is represented phonetically but could also be conveyed in textual form ; prosodic information, which carries additional non-linguistic meaning and includes emphasis on words or phrases within sentences, question or exclamation intonation patterns or the emotional state ; and personal identity which enables the listener to recognize the speaker. Successful conversion requires transformation of only speaker-specific features to reflect the identity of the target and the reconstruction of speech signal having these modified properties. Unfortunately, no decomposition currently is in existence that would be able to represent the three distinct types of information conveyed by speech as orthogonal subspaces of parameters. It is consequently not possible to parameterize human voice using a vector of features, which could then be manipulated

in isolation from phonetic and prosodic information. Rather, the three types of information are tangled together in a manner which makes their separation infeasible. It is possible, however, to describe speech using a model of the process of speech articulation. Such a model, representing perceptually relevant characteristics of speech, comprises feature extraction from speech (analysis) and speech reconstruction from features (synthesis) stages. The features can then be used to determine, with the use of machine learning methods, certain statistical correlations that depend on speaker identity, and create models that can subsequently be used in the transformation of speech parameters toward the target voice. We can thus distinguish two stages of operation : the adaptation stage whereby the conversion function is determined, and transformation stage, in which an existing function is applied to achieve actual conversion. The adaptation is performed for each source-target speaker pair and is based on speech samples from both speakers. Most commonly, parallel corpus approach is employed in which utterances having the same content but spoken by different individuals are used with proper time alignment to account for speaking rate differences between speakers. Training sets, built by extracting features from time aligned phonetic events, instantiate speaker-specific distrubutions of speech features in the phonetic parameter space. The effect of adaptation is the estimation of a regression model mapping one space to anoter. During the conversion, the model is used to transform features extracted from new speech material of the source speaker toward the target, and the converted speech is subsequently reconstructed.

There are many prospective applications for voice conversion ; by far of greatest importance is the ability to create new voices for text to speech synthesis cheaply. For this, two conditions must be met : firstly, the quality of transformed speech must reach the level attained by state of the art speech synthesisers, and secondly, the system must be trainable from a limited amount of speech data (smaller than the amount required by the synthesiser for this application to make sense). Currently, however, voice conversion still does not come up to quality expectations and good performance is obtained only when sizeable speech corpora are given. Most of the recent research in the area of voice conversion focusses on quality and data set reduction. Another long term aim is to remove the parallel corpus requirement, which is an obstacle to practical application, since acquisition of parallel samples and their alignment necessitate expert knowledge. Ability to build conversion models from arbitrary utterances would enable widespread adoption of the technique in entertainment industry and is a prerequisite to cross-lingual conversion. Another long term goal is the creation of systems operating in real time, i.e. able to directly transform speech as it is spoken to a microphone, opening new application areas. This particular goal is the interest of this work, in which the design of a new system for voice conversion is presented.

Section 2 discusses the basics of parametric speech representation. The articulation process and its models are briefly reviewed and the four models of the spectral envelope used in this work are introduced. Section 3 gives an overview of the original speech analysis-synthesis system and describes its application in

voice conversion, which is its primary purpose. Experiments and preliminary results are presented in Section 4. Finally, Section 5 summarizes and concludes the paper.

2 Speech representation in voice conversion

2.1 The articulation process

Modelling speech is essentially modelling the process of its production, and thus articulation must be referenced in any speech-oriented study. The theory of speech production is now well understood [1, 2]. In most languages, speech arises as a result of excitation of the speech organs in one of the three possible ways (or a combination of these) : by a periodic oscillation of the vocal folds forming the glottis, by a turbulent noise produced by the air passing across narrow constrictions, or by impulsive release of air pressure due to momentary closure of the vocal tract. The excitation propagates across the vocal tract acting as a resonant cavity, whose resonance characteristics are dependent on its geometry and vary with the movements of the jaw, lips, tongue and the soft palatum. These movements change the longitidunal and cross-sectional shape of the vocal tract, and also switch the passage of air between the oral and nasal cavities. In effect, the frequency spectrum of the original excitation is coloured by the instantaneous characteristics of the vocal tract acting as a filter, producing a variety of sounds that build the phonetic system of any language. Additionally, because of anatomical differences between individuals, the produced sounds differ in a systematic way and this difference, which is out of articulatory control, is what constitutes the speaker's voice.

Glottal excitation, also referred to as phonation, is of special importance. It produces voiced speech which is almost periodic and consequently has harmonic structure of the spectrum (see Fig. 3). The fundamental period of oscillation in voiced speech is the pitch period and its reciprocal, the pitch frequency, denoted F_0, is the fundamental frequency common to all harmonic peaks in the spectrum. Pitch does not affect phonetic value of speech sounds, however, it does contribute to the perception of speech. The frequency of oscillation of the glottis depends on individual dimensions and structure of the larynx and is typically twice higher in female than in male voices. Additionally, it is controlled by muscle tension and the subglottal pressure, and thus it can be consciously changed during ariculation in order to give the utterance a desirable prosodic pattern. Prosody includes word accent and sentence accent, and is achieved by means of modulation of pitch frequency, amplitude and vowel duration. Unlike the word accent which has a well-defined placement, sentence accent can be placed depending on the context and the intended meaning, for example, by putting emphasis on particular words or phrases within the sentence. Modulation of pitch is also used for proper intonation, for example, an elevated tone indicates exclamation, and a raising pitch at the end of a sentence distinguishes it as a question, as opposed to declarative sentences having lowering pitch pattern. However, the exact way

of effectuating accent and other prosodic cues is also speaker specific and thus contributes to the perception of individuality.

2.2 Speech modeling

In practical speech processing, modelling all aspects of speech production is not necessary and the following source-filter model [1] is sufficient to successfully represent articulation.

Fig. 1. The source-filter model of speech production

In the model of Fig. 1, the source represents the kind of excitation (voiced or unvoiced) and is responsible for the harmonic structure of spectrum contributing pitch in voiced speech. The excitation spectrum is always flat, i.e. it does not account for the actual spectral shape of the glottal or fricative excitation, which is combined with the vocal tract resonances, lip radiation and other effects in the form of a single filter responsible for the spectral envelope of the obtained output signal. The model assumes the excitation source and filter are indepentent and thus lends itself for practical use in speech coding, since estimation of instantaneous envelope and pitch is possible in short windows (typically 20–50 ms) in which the signal can be considered stationary. Purely voiced speech can be aproximated as the output of a filter excited by a periodic impulse train, whereas fricative sounds are obtained by using white gaussian noise as input. In practice, many sounds exhibit both harmonic and noisy character, since the glottal tone and fricative noise in a constriction above the larynx can coexist during articulation, and the addition of noise in voiced exitation has been found to produce more natural sounding synthetic speech [3]. Thus, the excitation should be modeled a mixture of harmonic and noise components and is paremeterized by their time varying amplitude envelope and the fundamental frequency. Since the harmonic to noise proportion is related to the current phonemic value, it should remain unchanged in voice conversion, whereas the fundamental should be adapted to match the target speaker.

In contrast to the pitch which is a well-defined scalar quantity, the spectral envelope, representing the colouring of sound or the *timbre* of voice, is not uniquely defined. The representation of the spectral shape is a key problem in voice conversion and several different models exist. In this work, three models and four representations were considered, as described below.

Linear predictive envelope model. The linear predictive model is based on an attempt to predict new speech sample as a linear combination of previous samples :

$$\tilde{s}_n = -\sum_{i=1}^{N} a_i s_{n-i}, \tag{1}$$

In general, it is not possible to predict the evolution of the signal and hence there remains a residual $r_n = s_n - \tilde{s}_n$, given by

$$r_n = s_n + \sum_{i=1}^{N} a_i s_{n-i} = \sum_{i=0}^{N} a_i s_{n-i}. \tag{2}$$

The signal r_n is then the excitation which perfectly reconstructs the signal:

$$s_n = -\sum_{i=1}^{N} a_i s_{n-i} + r_n. \tag{3}$$

The parameters $a_n, i = 1, \ldots, N$, where N is the order of prediction, are obtained from the criterion of minimum residual norm [4] and fully specify the model. However, due to the sensitivity of the characteristics of the synthesis filter (3) to these coefficients, it can easily become unstable. Two other representations are used instead: the log area ratios (LAR) [5] and the line spectral frequencies (LSF) [6]. While the same spectral model can be described equivalently in both domains, their interpolation properties differ.

Homomorphic deconvolution. The homomorphic theory [7] allows the estimation of the envelope as the slowly-varying component of the logarithmic spectrum treated as a regular signal. A basic result from control theory is that the discrete Fourier spectrum Y_m of the response of a filter with transfer function H_m to an input having frequency characteristics X_m is $Y_m = X_m \cdot H_m$. Representing the amplitude spectrum in logarithmic domain, we obtain and additive relation

$$\log|Y_m| = \log|X_m| + \log|H_m|, \tag{4}$$

and thus the slowly varying envelope can be obtained from the so-called "cepstrum", or the Fourier transform of $\log|Y_m|$:

$$C_k = \sum_{m=0}^{M-1} \log|Y_m| e^{-i2\pi \frac{km}{M}} \tag{5}$$

The low-frequency components represented by $C_k, k = 1, \ldots, N$ account for large-scale variation of the log spectrum. Higher coefficients represent spectral detail and C_0 corresponds to signal power.

Mel-frequency spectrum. The mel-frequency cepstral coefficients (MFCC) had been initially introduced in speech recognition, but the underlying envelope model lends itself for speech reconstruction and can thus also be considered in voice conversion. The representation is based on the mel scale which approximates the frequency sensitivity of human auditory perception. The scale is given by the function

$$x(f) = 2595 \cdot \log_{10}(1 + f/700\,\text{Hz})\,[\text{mel}] \tag{6}$$

and gradually changes shape from linear to logarithmic with growing frequency. The mel-frequency filter bank is then defined as a set of P filters uniformly spaced in the mel scale, corresponding to approximately equal sensitivity of the human ear. The filtering is performed in the frequency domain using triangular frequency weighting masks F_p shown in Fig. 2. Subband energy level E_p is calculated in each frequency band as a scalar product of the mask and the spectrum amplitude. From these levels, a piecewise-linear envelope model can be reconstructed. The envelope is then represented using N initial coefficients of the discrete cosine transform

$$C_k = \sum_{p=1}^{P} E_p \cos \frac{k(2p-1)\pi}{2P}. \tag{7}$$

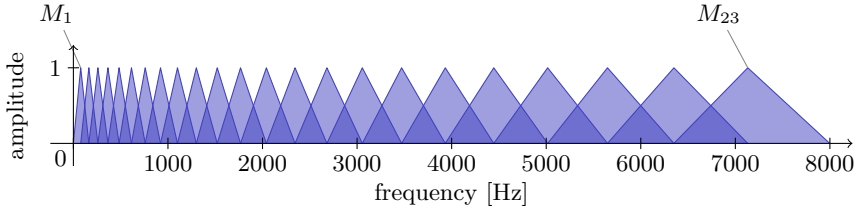

Fig. 2. The mel filterbank spectrum defined by overlapping triangular masks

3 Realtime speech transcoding

In this work, a mixed-excitation type of vocoder is used, designed with low latency and low computational cost as primary goals. The original idea presented in [8] is extended here with new envelope representations.

Similar to other mixed-excitation vocoders, speech is synthesised from a flat mixture of white noise and regular pulse train. However, it differs in the way the proportion is estimated from speech. Most vocoders are based on classification of speech frames into voiced or unvoiced and process each case differently. This leads to problems e.g. in estimating weakly voiced cases. The mixed excitation linear predictive (MELP) coder adressed this issue by allowing the voicing to vary from strongly and weekly voiced, by estimating the strength of harmonicity in the voiced case and changing the noise proportion in the mixture accordingly [3]. However, it still relies on voiced/unvoiced decision. A different approach is used in the harmonic-noise model (HNM) of [9], in which a maximum voiced frequency is estimated from speech, allowing voiced and unvoiced frames to be

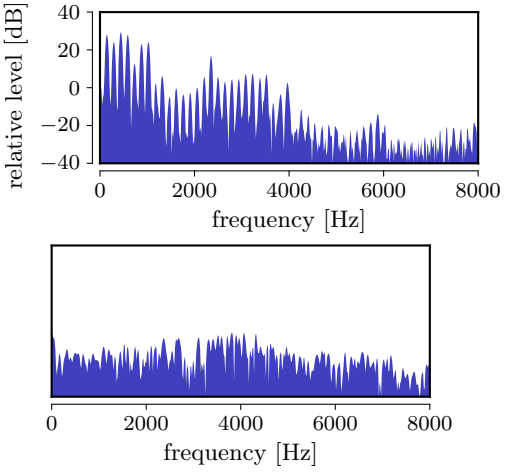

Fig. 3. Frequency spectrum of voiced speech (left, phoneme /o/) and unvoiced speech (right, phoneme /f/). Harmonic structure distinguishes voiced speech from unvoiced, which resembles coloured noise.

handled uniformly. However, in this approach, the excitation is not a weighted mixture of periodic pulse train and noise but obtained by combining the harmonic band below the estimated threshold and stochastic above. As a result, it cannot realistically reflect all kinds of phonation.

In an attempt to fill the gap between these vocoders, a new method of voicing estimation was introduced which allows the synthesis of excitation by mixing white gaussian noise and periodic pulse train in arbitrary proportion. The signal is then fed into the synthesis filter to shape the spectrum and give is the desired phonemic value and individual voice colouring. In order to be able to synthesise the excitation, it is necessary to estimate three parameters: the pitch frequency, the harmonic amplitude and the inharmonic (noise) amplitude.

Pitch tracking is done in autocorrelation domain. A distinguishing feature of the proposed coding scheme is that it does not rely on classification into voiced and unvoiced speech. Instead, a continuous degree of voicing is introduced which takes value from interval $[0, 1]$ and is interpreted as the ratio of harmonic to total energy. The estimation of this parameter, illustrated in Fig. 6, is based on the periodicity of autocorrelation sequence of harmonic (voiced speech) signals and the fact that the zeroth autocorrelation coefficient is the signal energy in the analysis frame. Consequently, the degree of voicing serves as the basis for estimation of the harmonic and noise energy levels from total frame energy, from which respective envelopes used in synthesis are generated. Since there is no voiced/unvoiced distinction, the fundamental frequency is estimated in all frames and is used for generating the periodic excitation. However, in unvoiced segments, its amplitude is small and thus the periodic component does not appreciably affect the nature of fricatives.

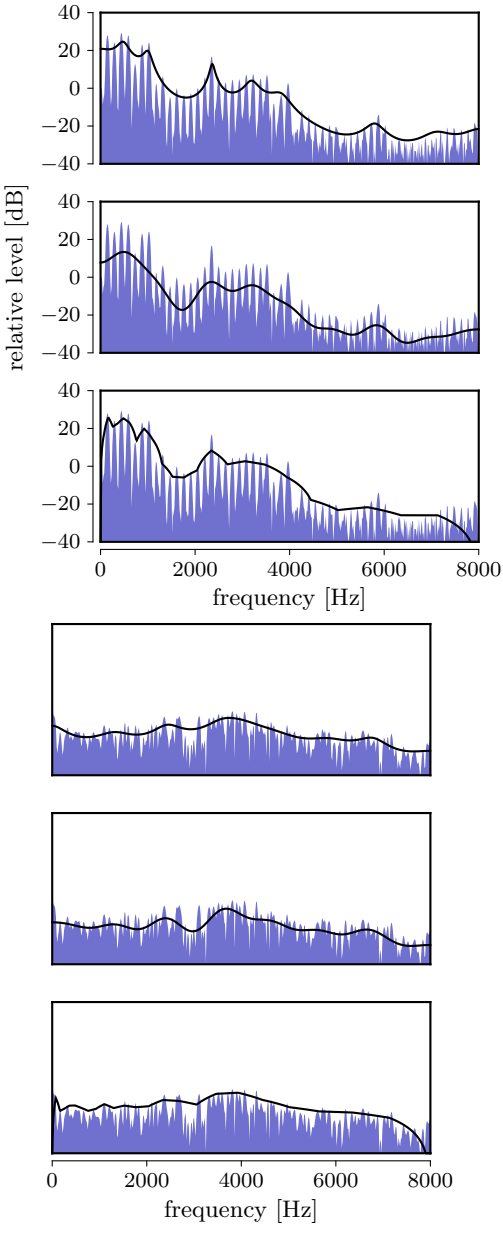

Fig. 4. Spectral envelope models for the same voiced (left) and unvoiced (right) speech spectra as in Fig. 3. From top to bottom : linear predictive (LPC) model, homomorphic model, mel-frequency cepstral model.

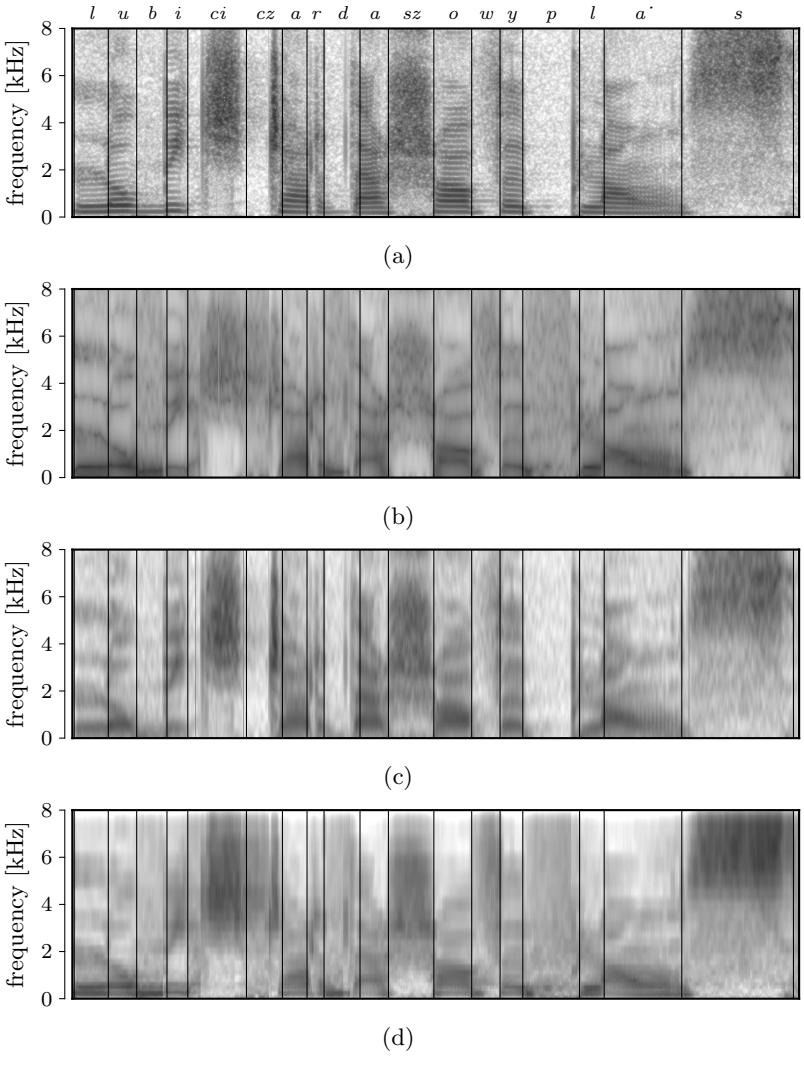

Fig. 5. Spectrograms of sample utterance „Lubić czardaszowy pląs" by female subject (a) and fitted spectral envelopes: linear predictive model (b), homomorphic model (c), mel-frequency cepstral model (d). All models are described using $N = 18$ parameters. Phonetic boundaries and labels taken from *CORPORA* annotation information.

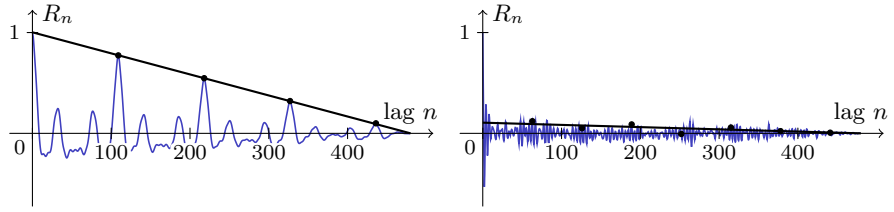

Fig. 6. Estimation of the degree of voicing by regression in autocorrelation domain. The degree of voicing is obtained as the intersection of the regression line (black), fitted to peak points (dots) of the autocorrelation function (blue), with $n = 0$ and has the interpretation of a ratio of harmonic to total energy. Left : voiced phoneme /o/, right : unvoiced phoneme /f/.

In voiced speech, the degree of voicing is not constant across frequency but typically decreases at higher frequencies because the envelope of glottal excitation falls off with frequency [1] (it can readily be seen in Fig. 3 as a decreasing dynamic range between harmonic peaks and valleys). To account for this, estimation of the degree of voicing in selected frequency bands is possible. In this work, four frequency bands are used (0–1 kHz, 1–2 kHz, 2–4 kHz, 4–8 kHz). Since the autocorrelation can be equivalently computed from the time and frequency domain, subband filtering can efficiently be performed by masking selected frequency regions of the spectrum. Fig. 7 illustrates the voicing estimation on real examples (which include stop /t/ and trill /r/ consonants), both in the full band and subband cases. It can be observed that both fricative and plosive phonemes cause the estimate to drop. When the energy envelope grows while the voicing is low, a noise impulse is generated in synthesis which is used to obtain plosive sounds. It can be observed that voicing also drops around phoneme transitions which is an undesirable effect. The subband estimates are more noisy, but as a rule the lower frequency bands have higher voicing than the high frequency bands in voiced segments which agrees with what is observed in the spectra.

The four different spectral shape representations (linear prediction represented in LAR or in LSF domain, homomorphic envelope and the mel-frequency cepstral envelope) have been built into the system as different modes of operation. A comparison of the different envelope models for a sample utterance is presented on Fig. 5. The spectrograms represent time-frequency pattern of a sample utterance and the corresponding spectral envelopes obtained using the different models.

Speech coding requires the estimation of the pitch period, the degree of voicing in four frequency bands and N envelope parameters (typically not exceeding 20 at 16 kHz sampling frequency) for a total number of up to 25 parameters per frame. The same window length of 30 ms is used for estimating envelope for each envelope model, as well as for pitch estimation. This translates into a minimum buffering delay of 30 ms. Envelope smoothing filter contributes an

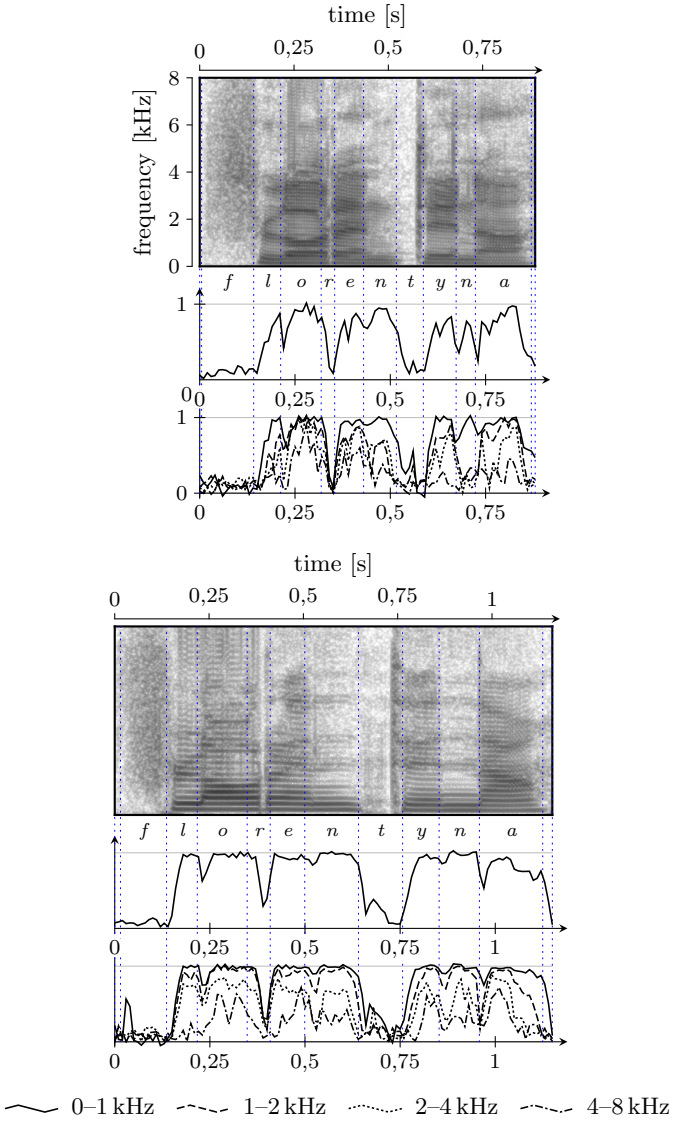

Fig. 7. Dynamic behaviour of voicing estimator for a sample utterance „Florentyna". From top to bottom : spectrogram image with phonetic segmentation and labels from *CORPORA* annotation files, fullband voicing estimate, subband voicing estimates in four frequency bands. Left : male voice, right : female voice.

additional delay of 10 ms and the synthesis filter delay can be neglected. As a result, conversion with a latency below 50 ms is possible.

The computational cost depends on the envelope model used. It is lowest in case of LAR, where parameters are estimated in the efficient Levinson-Durbin recursion. Its complexity is bounded by the square of the prediction order, which is typically not greater than 20. LSF require root finding, but because of special properties of the roots, efficient search procedures exist (see e.g. [10]). The highest cost is incurred by spectral representations (cepstral coefficients and MFCC) but their execution below real time on ordinary PC machines is not a problem.

4 Voice conversion experiments

The speech analysis-synthesis system described in the previous section served as a basis for building a voice conversion system. Speech data were taken from *CORPORA* parallel corpus of spoken Polish [11]. For adaptation phase, data sets were derived that were composed of the envelope parameters of selected kind and the pitch frequency profiles. The degrees of voicing were not transformed.

Pitch modification was based on a simple scaling which is a standard approach comonly found in voice conversion literature; in this method, the mean and variance of the source speaker's pitch are adjusted to that of the target speaker. This method preserves the prosody of the source utternace which may affect conversion performance, since pitch is known to carry some speaker-dependent information (e.g. learned manners of articulation). However, it is also known that when one speaker attempts to vocally impersonate another, the speaking style changes are the easiest to achieve. It is much harder to change the *timbre* of voice, since it is determined by anatomical conditions and thus spectral envelope transformation has traditionally received the greatest research focus.

For envelope modification, simple linear transformation is not adequate since the coupling between speaker's individuality and the phonetic content is generally nonlinear. The properties of voice reside in details of formant positions, bandwidths and envelopes in each phoneme. Through a careful time alignment of the data such that parameter vector pairs are obtained from analogous phonetic events, the phonetic variability in the data set is minimized, leaving mainly voice-induced differences, making it possible to obtain the appropriate transformation function by using machine learning methods. In this work, time alignment was based on phonetic annotation which accompanies the *CORPORA* corpus of speech. Three speakers were selected (a man, a woman and a boy) to cover both low and high-pitched voices. Only utterances with identical phonetic annotation across the selected pair of speakers were included into the data set. Additionally, the training data were generated from isolated single word utterances only, in order to minimize prosodic differences between speakers which can arise in longer utterances due to different understanding of the meaning of spoken sentences. Within each phonetic segment, ten equidistant frames were extracted along the segment and the corresponding frames were used to extract envelope param-

eters. The constructed training sets contained around 14000 vectors (varying depending on source-target pair).

After the construction of the training set, artificial neural network was trained using backpropagation algorithm. The network was a standard multilayer perceptron topology with bipolar sigmoidal (*tanh*) activation function in the hidden layer and linear output layer. The input and output layers were of size equal to the number of features, while the hidden layer size was set to 50 units. A validation subset was set aside to prevent the network from overfitting. An adaptively varying step size was used and the maximum number of iterations was 3000.

The trained neural network was subsequently used in actual conversion. In this mode, the same analysis procedure as during training set construction was applied. The envelope parameters were transformed using the neural network and the modified parameters were used to reconstruct the target speaker's envelope. The new filter was used with the excitation generated using transformed pitch to yield transformed speech samples.

The obtained samples were used in a formal listening test to evaluate the performance and quality of the method using subjective opinions of independent listeners. Performance was measured in ABX-type of test, in which the participants were presented two reference samples A and B of the same utterance from source and target speaker, in random order. Then the tested sample X was presented and the task was to decide if X was closer to A or to B. In the case the tested sample was judged closer to the target, a success was counted, and the percentage of successful speaker identity conversions was used as the perforance metric. A second test was the speech quality evaluation based on mean opinion score (MOS) test. A standard 5-point scale was used according to ITU recommendation [12] and the answers were averaged to give a quality score.

Table 1. ABX and MOS results for different spectral envelope representations and feature vector dimensions in regular voice conversion

	ABX		MOS	
feature dimension W	15	18	15	18
LAR	**80,0%**	62,5%	**2,53**	2,50
delta LSF	**70,0%**	66,7%	2,20	**2,42**
cepstral	66,7%	**75,0%**	1,93	**2,08**
MFCC	66,7%	**87,5%**	2,13	**2,42**

As a baseline for judging the quality of conversion, pure analysis/synthesis results (without parameter modification) were evaluated in a separate listening test representing the maximum achievable quality level. The average MOS scores obtained were 3,3 – 3,5 for LAR, 2,8 – 3,1 for MFCC, and 2,5 – 2,8 in LSF, depending on the feature vector dimension, demonstrating room for improvement.

5 Discussion

While ABX performance results indicate that the conversion was successful, the quality scores are too low for practical use. The quality degradation comes from two sources: firstly, parametric speech coders are always lossy and their quality is known to be lower compared to waveform coders at comparable baud rates. This is because waveform coders can represent also speech features which do not fit into models. For example, CELP coders (e.g. the GSM EFR) encode the most perceptually important part of LPC residual in a compact form, allowing to account for antiformants or transients, which are difficult to properly represent and estimate. On the other hand, in voice conversion, where baud rate reduction is not the aim, parameter quantization is not necessary and thus this source of distortion is eliminated. The quality achieved in the direct transcoding scenario (analysis followed by synthesis, with no feature transformation) is lower but still comparable to most speech coders used in the industry (for a survey see e.g. [13]) and could be acceptable for some applications. The main reason for quality loss is the limited expressive power of the excitation model, which cannot adequately represent some effects such as aperiodicity in voice onset. Underestimation of voicing around phoneme transitions also impacts quality, since it results in excessive noise in these time instants. By increasing the number of features, the quality could be improved within certain limits but that in turn would necessitate more training data for a reliable estimation of the conversion function. As a tradeoff, the dimension of envelope feature vectors should be confined to the range 15-20.

The second source of distortion is the transformation function learned from sample speech data. As can be seen from Table 1, the quality with conversion function is almost 1 MOS degree lower than the direct coding baseline, indicating a significant impact of the conversion function. The negative effect of the conversion function comes from several sources. One of them are the speaking style differences and other discrepencies in the training data, leading to errors in the conversion function. A remedy can be the creation of a dedicated corpus, where speaking style differences are minimized by the speakers following a reference prosodic pattern, as proposed by Kain and Macon [14]. This, however, would also necessitate additional annotation work since the training data are aligned based on phoneme boundaries; the special corpus requirement is also not practical and research is now aimed at learning from general non-parallel corpora. The second notable problem is the class imbalance due to unequal occurrence of phonemes, which may cause the less frequent ones to be underrepresented and suffer more distortion upon transformation.

A separate issue is the perceptual relevance of the particular representation of the spectral envelope. As can be seen from Fig. 4, the linear predictive model gives a very close approximation. However, detailed analysis revealed that interpolation of envelopes in both LAR and LSF domains has important shortcomings. Intuitively, a perceptually correct average of two tones of different frequency and amplitude should be a single tone having average frequency and average amplitude, not a mixture of two tones with half their original amplitude. In both

LSF and LAR representations, it was found that interpolation causes broadening of formant peaks, resulting in a muffled quality of reconstructed signal. While interpolation is not explicitly used in the conversion system, machine learning is inherently based on averaging over learning examples and thus causes loss of spectral detail. The homomorphic envelope model, on the other hand, can be seen to systematically underestimate the real envelope which due to the fact that it is a low-frequency fit to the spectrum and thus passes midway between peaks and valleys. As a result, this representation gave the worst results even in pure coding case. The mel-cepstral representation appears to be a promising alternative to LPC representations, however, MFCC did not outperform LAR representation in listening tests. Supposedly, the resolution of the filter bank is too high in the low frequency region, causing original pitch to affect the fitted model, which then interferes with the generated excitation with a modified pitch.

Acknowledgements

The study is cofounded by the European Union from resources of the European Social Fund. Project PO KL "Information technologies: Research and their interdisciplinary applications", Agreement UDA-POKL.04.01.01-00-051/10-00.

References

1. Fant, G.: Acoustic theory of speech production, with calcultions based on X-ray studies of russian articulations. I edn. Description and analysis of contemporary standard Russian. Mouton & co, The Hague (1960)
2. Flanagan, J.: Speech Analysis Synthesis and Perception. Kommunikation und Kybernetik in Einzeldarstellungen. Springer-Verlag Berlin Heidelberg (1965)
3. McCree, A., Barnwell, T.P., I.: A new mixed excitation LPC vocoder. In: Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on. (Apr 1991) 593–596 vol.1
4. Makhoul, J.: Linear prediction: A tutorial review. Proceedings of the IEEE **63**(4) (April 1975) 561–580
5. Viswanathan, R., Makhoul, J.: Quantization properties of transmission parameters in linear predictive systems. Acoustics, Speech and Signal Processing, IEEE Transactions on **23**(3) (Jun 1975) 309–321
6. Itakura, F.: Line spectrum representation of linear predictor coefficients of speech signals. The Journal of the Acoustical Society of America **57**(S1) (1975) S35–S35
7. Oppenheim, A.: Superposition in a class of nonlinear systems. PhD thesis, Massachusetts Institute of Technology (1965)
8. Lenarczyk, M.: Parametric speech coding framework for voice conversion based on mixed excitation model. In Sojka, P., Horák, A., Kopeček, I., Pala, K., eds.: Text, Speech and Dialogue. Volume 8655 of Lecture Notes in Computer Science. Springer International Publishing (2014) 507–514
9. Stylianou, Y.: Modeling speech based on harmonic plus noise models. In Chollet, G., Esposito, A., Faundez-Zanuy, M., Marinaro, M., eds.: Nonlinear Speech Modeling and Applications. Volume 3445 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2005) 244–260

10. Lenarczyk, M.: Robust and accurate lsf location with laguerre method. In: Interspeech 2015, Dresden, ISCA (September 2015) 423–427
11. Grocholewski, S.: Corpora - speech database for polish diphones. In: Proceedings of Interspeech, Rodos (1997) 1735–1738
12. ITU-T: Recommendation P.800, Methods for subjective determination of transmission quality (1996)
13. Ramo, A., Toukomaa, H.: On comparing speech quality of various narrow- and wideband speech codecs. In: Signal Processing and Its Applications, 2005. Proceedings of the Eighth International Symposium on. Volume 2. (August 2005) 603–606
14. Kain, A., Macon, M.: Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction. In: Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on. Volume 2. (2001) 813–816 vol.2