APPLYING IMPROVED SPECTRAL MODELING FOR HIGH QUALITY VOICE CONVERSION

Fernando Villavicencio

Music Technology Group
Universitat Pompeu Fabra
Ocata 1, 08003 Barcelona, Spain

Axel Röbel, Xavier Rodet

IRCAM-CNRS-STMS
Analysis-Synthesis Team
Place Igor Stravinsky, 75004 Paris, France

ABSTRACT

In this work, accurate spectral envelope estimation is applied to Voice Conversion in order to achieve High-Quality timbre conversion. True-Envelope based estimators allow model order selection leading to an adaptation of the spectral features to the characteristics of the speaker. Optimal residual signals can also be computed following a local adaptation of the model order in terms of the F_0. A new perceptual criteria is proposed to measure the impact of the spectral conversion error. The proposed envelope models show improved spectral conversion performance as well as increased converted-speech quality when compared to Linear Prediction.

Index Terms— Speech synthesis, speech analysis, cepstral analysis, spectral analysis, linear predictive coding.

1. INTRODUCTION

Interest in Voice Conversion (VC) has increased significantly in recent years on the speech technology comunity. A Voice Conversion system is conceived to process the speech signal in order to render the perceived identity of a source speaker into that of a target speaker. The expected qualities of the converted speech should be those of realistic speech synthesis (naturalness, intelligibility) and clearly, a succesful perception of the identity conversion.

In general, the VC systems based on statistical parameter conversion of voiced speech by Gaussian Mixture Models (GMM) show the best results. This method can be briefly described as a GMM-based modeling of the acoustical space of the timbre of a speaker. This information, represented by the spectral envelop, is commonly modeled by Linear Prediction (LP) and parametrized in the form of Line Spectral Frequencies.

Despite the extensive work carried out since the emergence of VC research, high-quality voice conversion has not yet been achieved. VC systems have been restricted to low sample-rate speech and the resulting performance is not considered satisfactory. Besides the remaining challenge to model and convert the speaking style, the quality of the converted timbre is not always found to be natural and artifacts free. While the latter effect is primarily related to the limited performance of the parameter conversion (i.e. the well-known oversmoothing effect by the use of statistical modeling [1]), it can also be due to factors such as the precision of the speaker-dependant features estimation and the capacity of the underlying synthesis framework to achieve a proper speech signal modification. Note that human perception has not been taken into account in the evaluation of the conversion performance.

The spectral envelope is defined by the smooth function passing through the prominent peaks of the spectrum. Accordingly, some envelope estimation errors can be translated as perceptive degradations when performing timbre modification. Moreover, significant estimation errors could degrade the source-target relationships defining the conversion function.

In our previous work, it was shown that precise envelope estimation can be achieved by means of True-Envelope based models [2],[3]. In the current paper we apply the improved estimation models to VC. Our baseline VC framework is based on the proposition found in [4]. An exhaustive experimental study was carried out aiming to evaluate the performance of the proposed methods and to compare them with LP in an objective way. The process was extended to higher sampling rate signals to achieve High-Quality Voice Conversion. One may expect in this way to outperform LP-based timbre conversion and therefore to improve converted speech quality. We will describe our experimental findings using optimal target information and perceptual cues when measuring the spectral conversion performance.

This article is organized as follows. Efficient spectral envelope estimation is introduced in section two. The improved spectral conversion is presented in section three. A perceptually-based evaluation framework of converted spectra is described in section four. The results of the experimental study comparing the spectral conversion performance by means of the different methods are found in section five. Finally, some conclusions are presented at section six.

2. EFFICIENT CEPSTRUM-BASED SPECTRAL ENVELOPE MODELING

The selection of the proper filter model (AR, MA, ARMA) and the corresponding model order are considered as main issues when seeking to perform efficient envelope estimation since they both are generally unknown. For speech signals, in particular, a physically motivated reasoning is usually considered to set the model order. Autoregressive and cepstrum based methods are commonly used as envelope models in the VC systems. The main problems observed with these methods come from the fact that they do not model the envelope in the way just defined. Moreover, the model order is rarely adapted to the characteristics of the speaker.

2.1. True-Envelope estimator

Efficient estimation can be achieved by means of the cepstrum-based True-Envelope (TE) estimator [5]. TE estimation is based on an iterative cepstral smoothing of the amplitude spectrum. The resulting estimation can be interpreted as the best band limited interpolation of the major spectral peaks.

2.2. Improved All-pole modeling

Aside the physical motivation to model voiced speech by an all-pole filter, another reason motivating autoregressive modeling on VC can be found on the good interpolation properties of the LSF parametrization. Nevertheless, as explained in [6], LP suffers from aliasing and it does not represent the desired spectral information passing through the prominent peaks. Usually, the resulting LP spectra will overestimate the predominant maximas of the spectrum.

The Discrete All-Pole model (DAP) [6] aims to solve the aliasing problem of LP. It shows, however, some problems related to filter stability and the necessity of harmonic-peaks tracking. We found, in addition, a lack of robustness in DAP when aiming to perform an adaptation of the model order to the harmonic information [7], as the one presented in the next section.

Aiming to reduce the model mistmach of LPC, in [3] we presented an autoregressive model called True-Envelope LPC (TELPC), which uses the spectral envelope estimations obtained from TE as a target spectrum for the autocorrelation matching criteria. This proposal follows the idea introduced in [8] to use interpolated spectrum information for all-pole modeling.

2.3. Optimal envelope extraction by order selection

A major advantage of cepstrum-based methods is that an estimate of the optimal cepstral order can be provided, considering that for voiced speech the harmonic excitation spectrum samples the vocal tract filter with a sampling rate given by the current F_0. Denoting F_s the signal sampling rate, the resulting order selection has the form [2]

$$\hat{O} = \frac{F_s}{2F_0} \tag{1}$$

Note that the real optimal order, that is the order that provides an envelope estimate with minimum error, depends on the specific properties of the envelope spectrum. Nevertheless, the order selection according to (1) is a reasonable choice for a wide range of situations and the resulting error has been found through experimentation to be rather close to the one obtained with the real optimal order [2].

3. IMPROVING SPECTRAL CONVERSION

We summarize the advantages of applying improved spectral modeling to the VC baseline in three aspects: increasing precision of the spectral envelope estimation, adaptation of the features dimensionality to the characteristics of the speaker and increased spectral-flatness of the residual signal. Some ideas leading us to claim these benefits are addressed in this section.

3.1. Speaker-dependant features adaptation

Since GMM-based spectral conversion is restricted to use input and output vectors with constant dimensionality a local selection of the envelope order according to (1) cannot be performed. We propose, however, to apply an adaptation of the speaker features. For TELPC modeling, the order selection can be locally applied on the internal TE estimator used to fit the AR model. On the other hand, if using TE modeling, considering only the average F_0 of the speaker should lead us to a more judicious selection of the global order compared to any other arbitrary selection. Modeling of speaker features can then be adapted in average to the characteristics of the speaker. Note that following the joint source-target framework proposed in [9], it is straightforward to use different dimensionality for the envelope parameters of source and target speakers to achieve the GMM-based linear regression. Following these adaptations we expect to reduce the mismatching between the envelope estimations used as speaker-dependent features at the conversion and synthesis stage and the real information contained in the spectra.

3.2. Optimal residual computation

As an alternative to use residual prediction techniques that may create unnatural connections of residual-segments, we use local order selection to improve the residual estimation. As result, we obtain residual signals with significantly increased spectral-peaks flatness. It was found, through experimentation, an important improvement on the synthesis quality when a modified envelope is applied in this way. We observed that increased spectral-flatness leads generally to a perceived reduction of features belonging to the original speaker.

4. EVALUATION OF THE SPECTRAL CONVERSION

4.1. GMM-based learning validation

The adaptation of the envelope model order to the mean F_0 of the speaker, combined with the use of speech signals with increased sampling rates will generally result in an increased dimensionality of the feature vectors. As a consequence, the typical values of data size and GMM components must be evaluated since it is well-known that statistical modeling suffers from the curse of dimensionality. In general, the number of feature vectors needed to achieve learning generalisation increases exponentially with the dimensionality. On the other hand, we do not expect a significant augmentation in the number of components since an increased precision of the envelope information should not modify, in principle, the average number of representative patterns in the acoustical space of the speaker (they are mostly defined by the overall shape of the spectrum). However, we must remark that the dimensions representing finer details of the envelope could be poorly correlated between both speakers leading to small values in the covariance matrixs. In such cases, we can deduce from the linear regression model that the conversion will be reduced to the consideration of the mean target values.

4.2. Evaluation of converted spectra using perceptive criteria

Evaluation of VC systems has commonly been limited to the estimation of a transformation ratio based on spectral distortion measures related to the envelopes designed by the target features. This strategy appears to be a good indicator of the conversion performance since it can tell us how close the converted features are to the target related to the distance measured before conversion. However, we can not deduce how perceptually revelant the remaining differences between the converted spectra and the target spectra will be.

We propose therefore a new evaluation framework where an efficient estimation of the target envelope is used as reference information and perceptual cues are considered in the conversion error measure. Efficient estimates of the target envelopes are obtained by application of the optimal order criterion to each analysis frame. Then, inspired by [10], we propose the application of a perceptive model to the envelope estimates before the quantification of the conversion error. The perceptive model includes the non-linear scaling of the frequency axis, a model of the middle-ear filter and the consideration of the frame energy in terms of loudness based on Zwicker's law [11]. The resulting perceptually-based representations are then matched to compute the average error considering the whole spectrum as well as the perceptual bands.

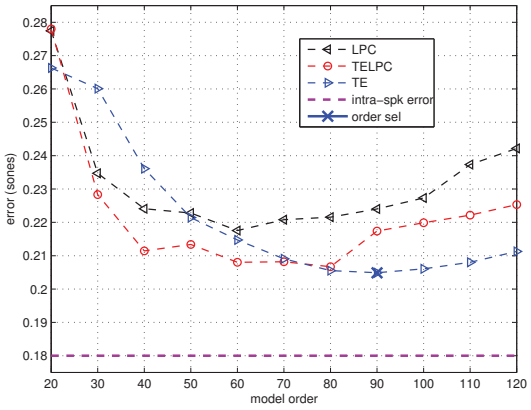

Fig. 1. Perceptual error as a function of the envelope model order. Male-to-male conversion.

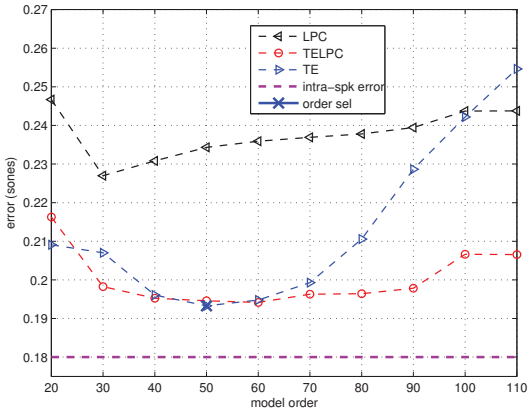

Fig. 2. Perceptual error as a function of the envelope model order. Female-to-female conversion.

The spectral conversion was carried out in an extensive experimental study aiming to compare the envelope models and to evaluate the proposed order adaptation. The resulting conversion error was measured using the perceptive model just described. The results obtained are presented in the next section.

4.3. Intra-speaker error

Acoustically speaking, the source-target matching of the envelopes can not be considered as optimal since it is obtained by Dynamic Time Warping from utterences containing different prosodic ($F0$) and duration properties for each speaker. Therefore, we can hardly expect a resulting conversion error lower than the variability produced by a single speaker when pronouncing the same phrase modifying prosody and duration. In our evaluation framework we considered the intra-speaker error experimentally measured from phrases pronounced in several ways by the same speaker. The resulting value can not be generalized since the proposed measure must to be evaluated among several speakers. However, this measure can help us to clarify how close the conversion error approaches to a natural intra-speaker variability level.

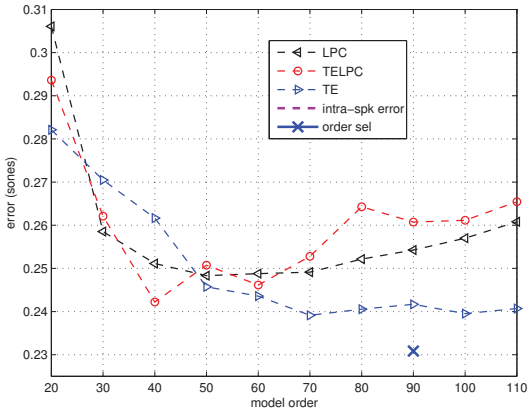

Fig. 3. Perceptual error as a function of the envelope model order. Female-to-male conversion.

5. EXPERIMENT AND RESULTS

5.1. Speech corpus and experimental framework

We built a higher quality speech database with sampling rate $F_s = 24Khz$ instead of the low-medium quality sampling rates typically used until now. This represents a challenge to the baseline system since it involves the extension of the probabilistic model to a frequency region where the correlation between the spectral information and the underlying classes captured by the model is theoretically weak. On the other hand, using higher quality signals should lead to a better understanding of the factors limiting the quality of the converted speech.

We used two male and two female speakers with mean F_0 values $M1 = 149Hz$, $M2 = 133Hz$, $F1 = 240Hz$ and $F2 = 238Hz$ in order to perform gender conversion and to better evaluate the effect of the proposed order adaptation based on the mean F_0 value. A total of two hundred short phrases considered to be phonetically balanced were recorded, providing a quantity of voiced features vectors within the range $[32,000 - 40,000]$ for the training set. Around $5,000$ vectors were used for the evaluation set (20 phrases). Source-target time-alignment was achieved by classical Dynamic Time Warping (DTW).

We performed the timbre conversion using an extensive grid of model orders covering the range $O = [20 - 120]$ with an increment of 10. This range is considered to be well adapated to our evaluation since it contains the order values typically found ($[20 - 30]$) and it largely covers the expected values when applying the order adaptation. We are also interested in supervising the learning behavior in order to clarify whether or not the increased dimensionality of the vectors affects the learning generalization. GMMs with $2, 4, 8$ and 16 components were considered. The spectral conversion was carried out using TE, TELPC and the commonly used method based on linear prediction (found as LPC on the figures) as envelope estimation methods.

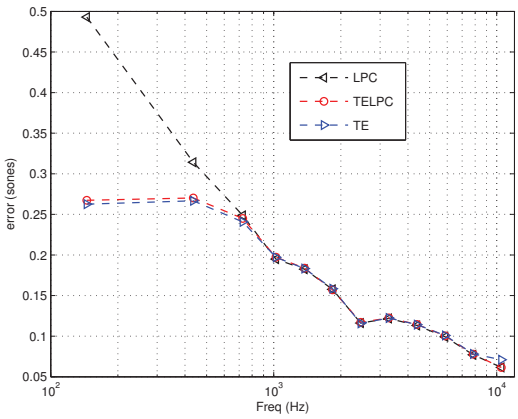

Fig. 4. Perceptual error as a function of the frequency band.

5.2. Results

In figures 1, 2 and 3 we show the results from the different gender spectral conversions. The curves correspond to the resulting perceptively-based conversion error measured in sones (loudness unit, [11]) where the model order is equally increased for both source and target speakers. The point denoted by 'x' represents the performance obtained when the average order adaptation is applied using the TE estimator for each speaker ($\hat{O}_{M1} = 80$, $\hat{O}_{M2} = 90$, $\hat{O}_{F1}, \hat{O}_{F2} = 50$). As mentioned above, TELPC uses local order selection to perform the TE estimate. The order that has been varied corresponds to the AR-model order used to approximate the TE estimate.

For the intra-gender cases (figs. 1 and 2), the order adaptation performs as well as the best fixed-order case. Note that for cases involving a female as target, the error curve shifts towards lower orders and its minimum is also found close to the case when order adaptation is applied. Also, note that in these female cases the performance of LPC decreases significantly which is expected due to the aliasing problem described in [6]. We found the female-male conversion especially interesting, where the proposed adaptation shows significant benefits. In this case, the source-target matching seems to take advantage of the adaptation of the envelope model to the charateristics of the speaker since the resulting error outperforms significantly the fixed-order cases. We remark that for this case, the order adaptation involves a conversion towards a higher dimensionality.

We can also appreciate that TELPC approaches faster to the minimum TE error than TE itself, allowing us to fix the order for both speakers without increase significantly the dimensionality of the feature vectors. This can be attributed to the benefits of performing local selection of the TE-order even if a low AR order is used.

Figure 4 shows an example of the resulting conversion error at each perceptual band (female-to-female conversion) when the model order corresponds to the order selection ($\hat{O} = 50$). Clearly, TE and TELPC significantly outperform LPC spectral conversion at the specially perceptually-important low frequency region. Also, when observing the learning performance as function of the features dimensionality, the overfitting effect was found rather low when increasing the order until the selection values. The best results were commonly obtained using 8 GMM components, affirming the values found in the bibliography.

To achieve converted speech synthesis, optimal residuals were used to perform PSOLA based source-filter synthesis. In cases involving gender conversion $F0$ normalization was applied. The overall results where affirmed by informal perceptive tests on the converted speech. The timbre of the synthesized speech using the order adaptation was perceived as more natural and clean than the one obtained through linear prediction. Some examples and results of this work are available in http://www.ircam.fr/anasyn/villavicencio.

6. CONCLUSIONS

We presented an improved spectral conversion framework for high-quality voice conversion based on efficient spectral envelope estimation. True-Envelope based estimators are used as estimation methods. GMM based VC was extended to higher quality speech processing. Spectral features dimensionality was individually adapted to the characteristics of the speaker and optimal residual signals were obtained by a efficient order selection. A perceptive evaluation framework was proposed to measure the perceptive impact of the conversion error. The resulting methodology outperforms LP based timbre conversion and shows increased converted speech quality.

7. REFERENCES

[1] A.W. Black T. Toda and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[2] A. Röbel, F. Villavicencio, and X. Rodet, "On cepstral and all-pole based spectral envelope modelling with unknown model order," *Elsevier, Pattern Recognition Letters*, vol. 28, no. 11, pp. 1343–1350, 2007.

[3] F. Villavicencio, A. Röbel, and X. Rodet, "Improving lpc spectral envelope extraction of voiced speech by true-envelope estimation," in *Proc. of ICASSP '06.*, France, 2006.

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

[5] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method," *Electronics and Communication (in Japanese)*, vol. 62, no. 4, pp. 10–17, 1979.

[6] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 411–423, 1991.

[7] F. Villavicencio, A. Röbel, and X. Rodet, "All-pole spectral envelope modelling with order selection for harmonic signals," in *IEEE ICASSP'07.*, 15-20 April 2007, vol. 1, pp. I–49–I–52.

[8] H. Fujisaki H. Hermansky and Y. Sato, "Spectral envelope sampling and interpolation in linear predictive analysis of speech," in *Proc. of ICASSP'84.*, 1984, pp. 2.2.1–2.2.4.

[9] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. of ICASSP'98.*, 1998, vol. 1, pp. 285–288.

[10] International Telecommunication Union, "Perceptual evaluation of speech quality (pesq)," Itu-t recommendation p.862, ITU-T.

[11] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, second updated edition. Springer, 1999.