

Analysis of LSF frame selection in voice conversion

Elina Helander¹, Jani Nurminen², Moncef Gabbouj¹

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

²Nokia Technology Platforms, Tampere, Finland

Abstract

In practical applications of voice conversion, it is necessary to be able to cope with small amounts of speaker-specific training data. Consequently, most of the proposed voice conversion algorithms are based on probabilistic conversion functions. Recently, however, there has been increased interest in unit selection based approaches for voice conversion. It is evident that typical training sets are too small for enabling meaningful selection of large units such as diphones. But would it be possible to use smaller segments like frames for high quality results provided that the selection is handled very well? In this paper, we analyze the performance of the frame selection approach in ideal conditions. In the experiments, line spectral frequencies of test sentences are replaced with the best matches from different training sets. The results show that perceptually transparent quality cannot be achieved with realistic database sizes.

1. Introduction

In unit selection speech synthesis [1], speech is produced by selecting segments from a recorded database and by concatenating them together. The database is large, typically consisting of several hours of speech, sometimes even tens of hours for providing an optimal unit sequence. The most popular unit sizes used in the selection are diphones and triphones. *Voice conversion* (VC) provides means for generating new text-to-speech (TTS) voices in a fast and easy manner using only small training sets. Voice conversion (or voice morphing) has inspired many researchers during the last two decades. The aim in VC is to convert speech from one speaker (source speaker) to sound like the speech from another particular speaker (target speaker).

Most voice conversion systems proposed in the literature are based on applying a conversion scheme directly to the source speech or its parametric representation. Typical examples of conversion schemes include Gaussian mixture model (GMM) based conversion [2] and the use of codebooks [3, 4]. Another approach for voice conversion is the parametric adaptation in a hidden Markov model (HMM) based speech synthesis framework [5]. All of the approaches share the same fundamental requirement: they have to be able to cope with small amounts of speaker-specific training data. Due to this requirement, the unit selection idea cannot be directly used with conventional unit sizes in voice conversion because there simply is not enough data to select from.

Speaker identity can be partially characterized using formant positions and bandwidths. Since it is very hard to handle the estimation of formants in a reliable and robust manner, the features most often used for conversion in VC systems are the line spectral frequencies (LSFs). LSFs are features that are derived through linear prediction (LP) where speech is modeled using a filter given by the LP coefficients and a residual. In most VC studies, the residuals are left unconverted, but there are strong arguments for converting residuals and some techniques have been proposed for this task for example in [6]. LSFs have also been used widely in speech coding, where typically large amounts of data from various speakers and languages is used for the training of LSF quantizers to obtain a good representation of the LSF space of all speakers. In speaker identification, high order LSFs have been reported to perform well as speaker identification features [7] and they have been used in many related studies (e.g. [8]).

Although LSFs seem to carry a lot of speech identity information, only a few personalized speech coding approaches have been proposed ([9, 10]).

The ultimate goal in voice conversion is to convert the speaker identity as accurately as possible while maintaining high speech quality. However, these requirements have been found to be somewhat contradictory in practice; better identity conversion usually requires more signal modifications that may cause more distortions. The main problem of the current VC techniques is that they are not very successful in changing the identity. Good results are mainly obtained because of forced ABX tests; the speech sample may sound more like target speech than source speech but it does not mean that it would ultimately sound like speech of the target speaker. All of the current techniques, including the GMM based conversion and the use of codebooks, have inherent drawbacks from this point of view.

Recently, Dutoit et al [11] proposed to first use a conventional GMM based approach to convert source LSFs to target LSFs and then search from the target speech database for the closest match to the converted LSFs in order to obtain more "realistic" target LSFs. The idea is attractive but can it help in achieving high quality conversion? In this study, we analyze if it is possible to select LSFs from a target database of a realistic size in such a manner that the quality of the converted speech would be very high or even indistinguishable from the target speech. The results of our experiments reveal how accurately LSFs could be chosen provided that the conversion is successful. Multiple speakers, different test sentence sets and different sizes of target databases are examined and the results are presented in the light of quality criteria used widely in speech coding.

This paper is organized as follows. In Chapter 2, the basic properties of LSFs and the related distance metrics and quality criteria are discussed. The experiments and results demonstrating the idealized frame selection performance are described in Chapter 3. Chapter 4 provides a short discussion on the results and Chapter 5 concludes the study.

2. Linear prediction and line spectral frequencies

Linear prediction is one of the basic techniques used in speech processing. This source-filter model can be used for separating a speech signal into linear prediction coefficients that model the vocal tract contribution and into an excitation signal. More precisely, the excitation signal, also referred to as the residual signal, can be obtained through LP analysis filtering,

$$r(t) = x(t) - \sum_{k=1}^m a_k x(t-k), \quad (1)$$

where $x(t)$ is the input speech signal and m is the order of the analysis filter $A(z)$. The linear prediction coefficients $\{a_k\}$ are usually estimated in a frame-wise manner using either the autocorrelation or covariance methods. The autocorrelation method is widely used because it always ensures that the resulting filters are stable.

For further processing, the linear prediction coefficients are often converted into the line spectral frequency representation. The fully reversible conversion can be carried out by first calculating the roots of the polynomials

$$\begin{aligned} P(z) &= A(z) + z^{-(m+1)} A(z^{-1}), \\ Q(z) &= A(z) - z^{-(m+1)} A(z^{-1}). \end{aligned} \quad (2)$$

Then, the LSF representation is formed simply by the angular positions $\{\omega_k\}$ of the complex roots in ascending order. The LSF representation is favored in different areas of speech processing for many reasons. For example, this representation offers advantageous properties from the viewpoint of quantization, interpolation and other processing, and it can guarantee filter stability.

The LSF representation has also been widely used in voice conversion. In selection based voice conversion, some distance measure is needed. The distance between two LSF vectors can be computed e.g. using weighted squared error with a diagonal weighting matrix,

$$d(\boldsymbol{\omega}, \hat{\boldsymbol{\omega}}) = (\boldsymbol{\omega} - \hat{\boldsymbol{\omega}})^T \mathbf{W}(\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}) = \sum_{k=1}^m w_k (\omega_k - \hat{\omega}_k)^2. \quad (3)$$

The weights can be used for approximating the properties of human hearing. We use the weights given in [13] defined by

$$w_k = c_k \left| H(e^{j2\pi f_k / f_s}) \right|^{0.6}, \quad (4)$$

where f_k denotes the frequency of the k th LSF element, f_s is the sampling frequency, and $H(z)$ denotes the synthesis filter $H(z) = 1/A(z)$. Furthermore, when dealing with 10-dimensional LSF vectors at a sampling frequency of 8 kHz, c_k is set to one for all k except for $c_9 = 0.64$ and $c_{10} = 0.16$, as proposed in [13].

In addition to the weighted squared error distance, another useful and popular metric for measuring the distance between two LP spectra is spectral distortion (SD). It is defined in dB as

$$SD = \sqrt{\frac{1}{f_u - f_l} \int_{f_l}^{f_u} \left(20 \log_{10} \frac{|H(e^{j2\pi f / f_s})|}{|\hat{H}(e^{j2\pi f / f_s})|} \right)^2 df}, \quad (5)$$

where f_l and f_u denote the lower and upper frequency limits of the integration. A convenient property of this measure is the fact that there are generally accepted SD based criteria for perceptual spectral transparency, i.e. criteria that guarantee that two spectra are indistinguishable through listening. In [13], it was concluded that transparency is achieved if the following three criteria are met: 1) average SD is less than 1 dB, 2) there are no outlier frames having SD above 4 dB, and 3) less than 2% of frames have SD in the range from 2 to 4 dB.

3. Experimental results

To study the performance level achievable in voice conversion using the frame-based selection approach, we carried out experiments in idealized conditions. The main idea in these experiments was to focus only on the frame selection by making the assumption that other parts of voice conversion would perform perfectly. In practice, we achieved this perfect conversion using recorded sentences from the target speaker as "converted test sentences". Frame-based selection was then applied on these recorded test sentences by replacing the LSF vectors in the test sentences with the best matches found in a selection database. The selection database was formed using uncompressed LSF vectors estimated from the speech of the target speaker. We experimented with various selection database sizes but different sentences were always used in testing and training, making the experiment realistic apart from the above-mentioned assumption of idealized conditions. Thus, the results achieved in these experiments demonstrate the upper bound for the performance of frame-based selection in voice conversion.

The experiments were carried out using the publicly available CMU Arctic database [12], a database of 1132 utterances spoken by 7 different speakers, 2 female and 2 male American English speakers, 1 Canadian English male, 1 Scottish English male, and 1 male speaker with Indian accent. The waveforms in the databases were downsampled to 8 kHz and 10th order LP analysis was performed at 10-ms intervals with overlapping 25-ms analysis frames, using the analysis module of the voice conversion system presented in [14]. Each analysis frame was windowed using a Hamming window and the LP coefficients were computed using the autocorrelation method.

Each speaker served as a reference speaker (speaker in test sentences) and as a selection database speaker for him/herself. In addition, each speaker was also used as a database speaker for the other speakers for comparison purposes. The number of sentences in the selection databases was varied (5, 10, 20, 50 and 100) by including new sentences in such a way that larger sets always also contained the sentences included in the smaller sets. All 80 reference sentences and the database sentence sets were selected randomly but they were kept the same for all speaker combinations.

The new LSFs replacing the LSFs in the original reference sentences were selected from the selection database using the weighted squared error distortion in Eq. (3) together with the weighting in Eq. (4). This scheme was used to obtain a reasonable computational complexity. The final results were evaluated using the spectral distortion formula in Eq. (5) since it provides the best comparison capabilities. Frames classified as silence were not included in the results. The average spectral distortion, measured in the range from 0 to 3.2 kHz, and the percentage of 2 and 4 dB outliers were calculated for two different categories: i) the reference speaker is the same as the database speaker (7 cases) and ii) the reference speaker is different than the database speaker (42 cases).

The mean SD averaged over all speakers is shown in Figure 1 for categories i) (solid line) and ii) (dash-dotted line). The best and worst results in category i) are also shown (dashed lines). The dotted line represents the mean values of each reference speaker's best results when selecting from another speaker's database, i.e. the result with another speaker's database that gave the lowest average spectral distortion values for the reference speaker. The mean percentage of 2 dB and 4 dB outliers is shown in Figure 2 and Figure 3, respectively.

As can be seen from Figure 1, the best matching LSFs were on average far away from ideal transparent quality. There are large differences between the speakers but even the best results were not very good. The low number of 4 dB outliers with larger databases is encouraging, but the requirement of having less than 2% of 2 dB outliers is far from being fulfilled. As expected, using other speaker's database was not as successful as using the speaker's own database, indicating that there are strong speaker-dependencies in LSFs. An interesting observation not directly visible in the figures was that the best results with other speaker's LSFs were always achieved when the LSFs were selected from a speaker with a matching gender. This is in line with the fact that the formant frequencies of female speakers are generally higher than the formant frequencies of male speakers due to the shorter vocal tract.

We also examined whether the quality would be much better if the number of sentences in the database was significantly increased. A set of 250 sentences resulted in an average SD of 1.3 dB for category i) with 7% and 0.1% of 2 dB and 4 dB outliers, respectively. The best result among the speakers was 1.15 dB. In addition, we tested if the usage of the whole Arctic database (1132 sentences minus one reference sentence) as the selection database could result in low spectral distortion. The mean of averaged SD was 1.09 dB for all speakers and the best speaker obtained an average SD of 0.97 dB, measured using 20 different reference sentences. There were 2.2 % of 2 dB outliers and no 4 dB outliers. Using the whole database offers almost transparent quality. For the best other speaker, the average SD was 1.45 and the percentage of 2 dB outliers about 14%. Nevertheless, the database of this size would not be suitable for practical voice conversion.

4. Discussion

LSF selection from a single frame does not seem to provide very high spectral quality if the size of the database is realistic from the viewpoint of practical applications. In [11], the authors do mention that there is a relatively large non-parallel database available – which means in their case over 12 minutes of data. It can be considered as a very large database for voice conversion. This would equal to almost 250 sentences if a sentence is on average 3 seconds long. The results presented in this paper show that transparent quality cannot be achieved even with this kind of relatively large database in idealized conditions.

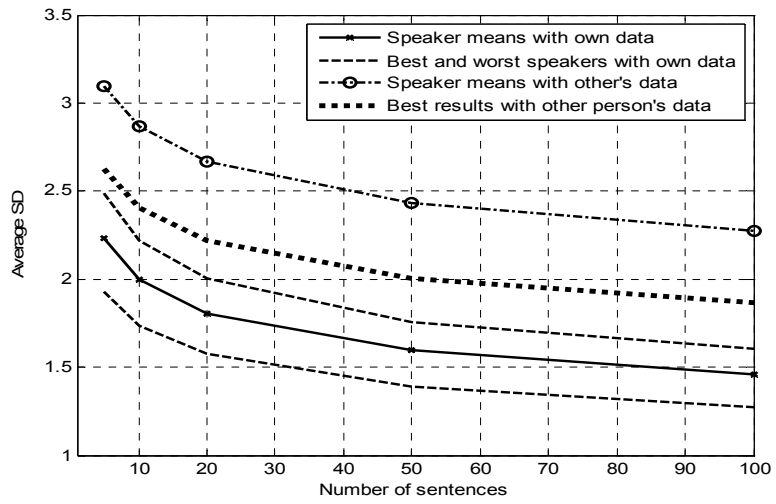


Figure 1. Spectral distortions of LSF databases gathered from the same speaker or from other speakers.

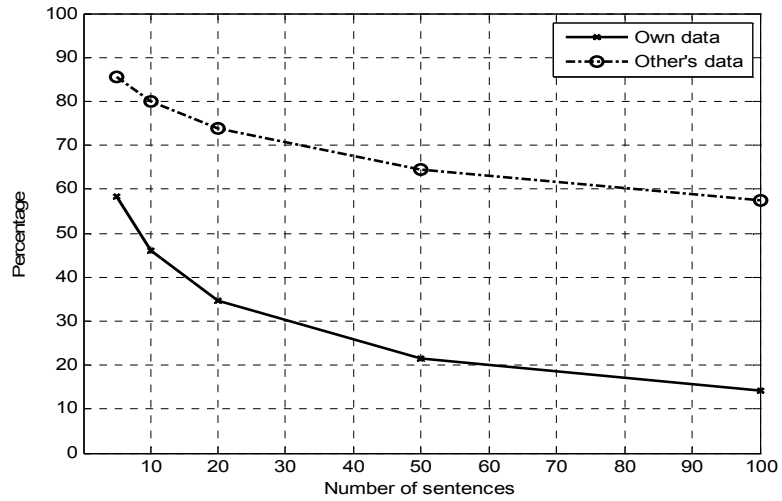


Figure 2. The mean percentage of 2 dB outliers for all speakers

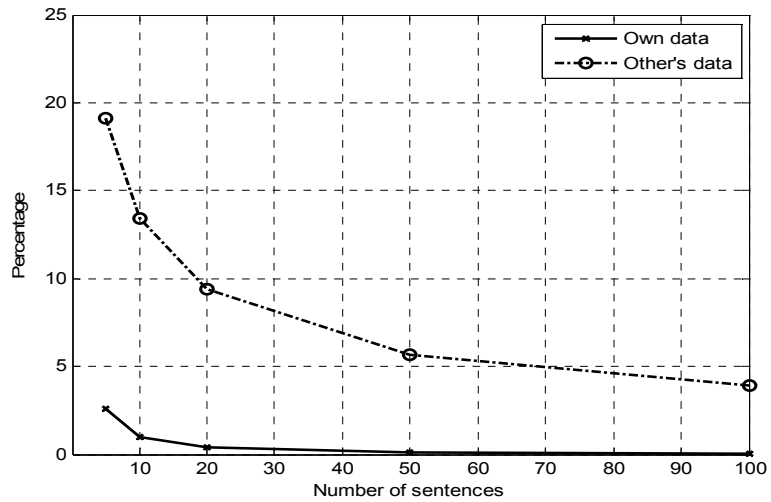


Figure 3. The mean percentage of 4 dB outliers for all speakers.

Even if there would be enough sentences to fulfill the requirement of transparent or otherwise very high quality, there is no target signal available during the conversion and thus the selections must be based on the source speaker's sentence. This moves the realistically achievable quality even further away from the transparent level. Moreover, we have only considered LSFs in this study. In reality, there is also a need for transforming the residual. Residual selection techniques have been proposed to be based on the LSF vector and its corresponding residual. In [6], residual selection was analyzed and it was found that the selection of an optimal LSF sequence similarly as in unit selection can be more preferable than direct selection without considering neighboring frames. Nevertheless, the residual selection was ultimately based on the converted LSF vector, and it is reasonable to assume that residual selection will be even more challenging than LSF selection.

5. Conclusions

In this paper, we analyzed whether it is possible to select LSF vectors from a small database with very high quality in the scope of voice conversion. The CMU Arctic database with 7 speakers was used to test if a small set of sentences could act as an effective selection database in a voice conversion. We found that small database sizes commonly used in voice conversion are not adequate for representing the LSF space of a speaker and the achievable quality is far from transparent quality even in ideal conditions.

References

1. *A. Black, A. Hunt*. Unit selection in a concatenative speech synthesis system using a large speech database. In Proc. of ICASSP, pp. 373-376, 1996.
2. *Y. Stylianou, O. Cappe, E. Moulines*. Continuous probabilistic transform for voice conversion. IEEE Trans. on Speech and Audio Processing, vol. 6(2), pp. 131-142, March 1998.
3. *M. Abe, S. Nakamura, K. Shikano, H. Kuwabara*. Voice conversion through vector quantization. In Proc. of ICASSP, pp. 565-568, 1988.
4. *O. Turk, L. M. Arslan*. Robust processing techniques for voice conversion. Computer Speech and Language, vol. 4(20), pp. 441-487, October 2006.
5. *J. Yamagishi, K. Ogata, Y. Nakano, J. Isogai, T. Kobayashi*. HSMM-based model adaptation algorithms for average voice-based speech synthesis. In Proc. of ICASSP, vol. I., pp. 77-80, 2006.
6. *D. Sündermann, H. Höge, A. Bonafonte, H. Ney, A. Black*. Residual prediction based on unit selection. In Proc. of ASRU, pp. 369-374, 2005.
7. *D. Reynolds*. Experimental evaluation of features for robust speaker identification, IEEE Trans. on Speech and Audio Processing, Vol. 2, no. 4, pp. 639-643, October 1994.
8. *T. Kinnunen, E. Karpov, P. Fränti*. Real-time speaker identification and verification, IEEE Trans. on Audio, Speech and Language Processing, vol. 14, no. 1, pp. 277-288, January 2006.
9. *W. Jia, W.-Y. Chan*. An experimental assessment of personal speech coding. Speech Communication, vol. 30, no. 1, pp. 1-8, 2000.
10. *C.-H. Lee, S.-K. Jung, H.-G. Kang*. Applying a speaker-dependent speech compression technique to concatenative TTS synthesizers. IEEE Trans. on Audio, Speech and Language Processing, vol. 15, no. 2, pp. 632-640, February 2007.
11. *T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Pérez, Y. Stylianou*. Towards a voice conversion system based on frame selection, in Proc. of ICASSP, vol. 4, pp. 513-516, 2007.
12. *J. Kominek, A. Black*. CMU Arctic databases for speech synthesis version 0.95. Technical report, Carnegie Mellon University, 2003.
13. *K. Paliwal, B. Atal*. Efficient vector quantization of LPC parameters at 24 bits/frame. IEEE Trans on Speech and Audio Processing, Vol. 1, no. 1, pp. 3-14, January 1993.
14. *J. Nurminen, V. Popa, J. Tian, Y. Tang, I. Kiss*. A parametric approach for voice conversion. In Proc. Workshop on Speech-To-Speech Translation, pp. 225-229, 2006.