LSF MAPPING FOR VOICE CONVERSION WITH VERY SMALL TRAINING SETS

Elina Helander[1], *Jani Nurminen*[2] *and Moncef Gabbouj*[1]

[1]Institute of Signal Processing, Tampere University of Technology, Finland
[2]Nokia Technology Platforms, Tampere, Finland

elina.helander@tut.fi, jani.k.nurminen@nokia.com, moncef.gabbouj@tut.fi

ABSTRACT

To make voice conversion usable in practical applications, the number of training sentences should be minimized. With traditional Gaussian mixture model (GMM) based techniques small training sets lead to over-fitting and estimation problems. We propose a new approach for mapping line spectral frequencies (LSFs) representing the vocal tract. The idea is based on inherent intra-frame correlations of LSFs. For each target LSF, a separate GMM is used and only the source and target LSF elements best correlating with the current LSF are used in training. The proposed method is evaluated both objectively and in listening tests, and it is shown that the method outperforms the conventional GMM approach especially with very small training sets.

Index Terms— voice conversion, line spectral frequencies

1. INTRODUCTION

Voice conversion (VC) is a relatively new field of speech signal processing. It aims at modifying speech spoken by one speaker (*source*) to give an impression that it was spoken by another specific speaker (*target*). The voice conversion process consists of two phases: training and conversion.

In training, a mapping from source features to target features is created based on training data from both speakers. In the conversion phase, any unknown utterance from the source speaker can be converted to sound like the target speaker. The most popular methods for creating the mapping in voice conversion include codebooks [1], [2], and Gaussian mixture models (GMMs) [3], [4]. Recently, voice conversion in the framework of hidden Markov model based speech synthesis has also become a popular topic (e.g. [5]).

GMMs have been found to offer reasonably good performance in voice conversion. On the other hand, the main drawbacks are over-smoothing and over-fitting. In [6], the over-fitting properties of different GMM based approaches were compared with different amounts of training data. It was concluded that the number of mixtures must be decreased when the amount of training data decreases.

Speaker identity can be partially characterized using formant positions and bandwidths. The estimation of formants is, however, difficult. The most common features used in voice conversion are based on direct use of spectral bands or on the source-filter theory. Examples of such features include MFCCs (Mel frequency cepstral coefficients) [3] and LSFs (line spectral frequencies) [4].

The most potential application areas of VC are related to entertainment. One of these applications is text-to-speech (TTS) voice customization. Usually new TTS voices are created using hours of recorded speech but voice conversion offers a way to generate new voices with much shorter recordings, typically in the order of 50-200 sentences. Nonetheless, it is still burdensome for the user; it is quite likely that a typical user is not willing to record even 50 predefined sentences for enabling his/her voice to be used as a TTS voice. A reasonable number of sentences would probably be in the order of 1-5 sentences but the current VC techniques cannot cope effectively with such training set sizes. One recently proposed idea for this problem is the use of eigenvoices [7], however, many-to-one conversion requires multiple pre-stored source speakers.

In this paper, we propose an approach for LSF conversion for the case where only a few training sentences are available. The approach is based on the joint density of the target and the source features, following the idea of [4]. However, the properties of LSFs are taken into account and the models are built for each target LSF separately based on the correlation coefficients of the joint source-target LSF vectors. This may reduce the model size and makes the GMM training more reliable when only small training sets are available.

This paper is organized as follows. LSFs and their properties as voice conversion features are considered in Section 2. In Section 3, we describe the proposed method of using a separate GMM for each LSF. Section 4 summarizes objective and listening test results when comparing the proposed approach against the conventional full-vector GMM approach. Some interesting findings are discussed in Section 5, while Section 6 concludes the paper.

2. LINE SPECTRAL FREQUENCIES IN VOICE CONVERSION

Line spectral frequencies have been widely used in many areas of speech processing and they have been particularly popular in speech coding. LSFs offer an alternative and fully reversible representation for linear prediction coefficients (LPCs). The conversion to line spectral frequencies is carried out by first forming the polynomials $P(z) = A(z)+z^{-(m+1)}$ and $Q(z) = A(z)-z^{-(m+1)}$ from the linear prediction analysis filter $A(z)$ of order m. The LSF representation is then formed simply by the angular positions $\{\omega_k\}$ of the complex roots in ascending order. The LSF representation offers many advantageous properties, for example in interpolation and quantization. Another significant benefit is that the use of LSFs guarantees filter stability. LSFs have also been used in many voice conversion systems, for example in [2] and [4]. One reason for the popularity in voice conversion is the close relationship to the modeling of the vocal tract and formants. Despite the highly beneficial properties, the use of LSFs as features in voice conversion also introduces some problems. For example, the k^{th} LSF coefficient may not always correspond to the same formant. Moreover, due to the ordering property, there are significant correlations between the LSF elements in a

This work was partially supported by the Academy of Finland, project No. 5213462 (Finnish Centre of Excellence program 2006 - 2011).

Table 1. The correlation coefficients for joint source-target LSF vectors.

	Source LSFs										Target LSFs									
	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
1	1.00	0.64	0.28	0.23	-0.13	0.04	0.00	-0.00	0.17	-0.21	0.54	0.59	0.29	0.07	-0.09	-0.21	-0.26	-0.10	-0.07	0.01
2	0.64	1.00	0.79	0.46	0.05	0.24	0.19	0.24	0.28	-0.01	0.39	0.80	0.75	0.55	0.21	0.05	0.02	0.17	0.23	0.17
3	0.28	0.79	1.00	0.58	0.18	0.37	0.26	0.40	0.26	0.17	0.21	0.66	0.84	0.82	0.44	0.27	0.31	0.34	0.44	0.34
4	0.23	0.46	0.58	1.00	0.63	0.49	0.32	0.33	0.40	0.13	0.18	0.42	0.44	0.45	0.63	0.55	0.36	0.26	0.23	0.11
5	-0.13	0.05	0.18	0.63	1.00	0.63	0.40	0.32	0.28	0.30	-0.05	0.03	0.10	0.17	0.59	0.84	0.63	0.37	0.24	-0.08
6	0.04	0.24	0.37	0.49	0.63	1.00	0.60	0.59	0.43	0.24	0.03	0.21	0.26	0.32	0.38	0.53	0.61	0.54	0.42	0.15
7	0.00	0.19	0.26	0.32	0.40	0.60	1.00	0.66	0.55	0.43	0.01	0.13	0.21	0.19	0.27	0.37	0.41	0.71	0.50	-0.10
8	-0.00	0.24	0.40	0.33	0.32	0.59	0.66	1.00	0.55	0.43	0.01	0.22	0.32	0.37	0.27	0.31	0.44	0.52	0.53	0.19
9	0.17	0.28	0.26	0.40	0.28	0.43	0.55	0.55	1.00	0.49	0.12	0.28	0.24	0.20	0.25	0.23	0.21	0.32	0.25	0.06
10	-0.21	-0.01	0.17	0.13	0.30	0.24	0.43	0.43	0.49	1.00	-0.13	-0.01	0.13	0.23	0.33	0.37	0.37	0.42	0.30	0.01
1	0.54	0.39	0.21	0.18	-0.05	0.03	0.01	0.01	0.12	-0.13	1.00	0.60	0.25	0.14	0.02	-0.12	-0.17	-0.07	-0.07	0.01
2	0.59	0.80	0.66	0.42	0.03	0.21	0.13	0.22	0.28	-0.01	0.60	1.00	0.75	0.54	0.22	0.02	-0.03	0.09	0.16	0.18
3	0.29	0.75	0.84	0.44	0.10	0.26	0.21	0.32	0.24	0.13	0.25	0.75	1.00	0.82	0.37	0.21	0.19	0.28	0.39	0.26
4	0.07	0.55	0.82	0.45	0.17	0.32	0.19	0.37	0.20	0.23	0.14	0.54	0.82	1.00	0.56	0.35	0.44	0.37	0.47	0.39
5	-0.09	0.21	0.44	0.63	0.59	0.38	0.27	0.27	0.25	0.33	0.02	0.22	0.37	0.56	1.00	0.77	0.56	0.37	0.27	0.13
6	-0.21	0.05	0.27	0.55	0.84	0.53	0.37	0.31	0.23	0.37	-0.12	0.02	0.21	0.35	0.77	1.00	0.76	0.47	0.34	-0.03
7	-0.26	0.02	0.31	0.36	0.63	0.61	0.41	0.44	0.21	0.37	-0.17	-0.03	0.19	0.44	0.56	0.76	1.00	0.66	0.51	0.22
8	-0.10	0.17	0.34	0.26	0.37	0.54	0.71	0.52	0.32	0.42	-0.07	0.09	0.28	0.37	0.37	0.47	0.66	1.00	0.70	0.04
9	-0.07	0.23	0.44	0.23	0.24	0.42	0.50	0.53	0.25	0.30	-0.07	0.16	0.39	0.47	0.27	0.34	0.51	0.70	1.00	0.31
10	0.01	0.17	0.34	0.11	-0.08	0.15	-0.10	0.19	0.06	0.01	0.01	0.18	0.26	0.39	0.13	-0.03	0.22	0.04	0.31	1.00

frame. This leads to the fact that diagonal covariance matrices do not work well in GMM based conversion. Regardless of these problems, LSFs can be considered good features for voice conversion.

Despite the fact that there are strong correlations between different LSFs, all the elements of the vectors are not correlated as can be seen from the correlation coefficient matrix of Table 1 calculated from CMU Arctic database [8] speakers *bdl* and *slt*. The correlation coefficients were calculated from joint source-target LSF vectors obtained using dynamic time warping (DTW) based alignment. The LSFs for the source and target were derived from 10th order LPCs estimated at 10-ms intervals using a 25-ms Hamming window and the autocorrelation method. The first ten rows and columns correspond to the intra-frame correlations in the source side whereas the last ten rows and columns correspond to the correlations in the target side. It is easy to see that, for example, the first and the tenth LSFs do not have a meaningful relationship in terms of correlation. In speech coding, this has given justification for the use of split vector quantization where the LSF vector is split into two or more parts that are quantized separately (e.g. [9]). The same idea is used in this paper but the motivation is coming from a different angle.

An interesting property resulting from the use of split LSF vectors is that less data is needed for occupying the LSF feature space than with full LSF vectors. In [10], it was experimentally verified that clearly more than 100 sentences are needed to cover the LSF space of one speaker in perceptually transparent manner with full LSF vectors. CMU Arctic database [8] with seven speakers and various sentence sets were used. We carried out a similar test with split LSFs following the 3-3-4 splitting proposed in [9] and the comparative results are shown in Table 2. According to the general criteria of perceptual transparency proposed in [9], it can be seen that only 5 sentences is enough for covering the LSF space with split LSF vectors if a few 4 dB outliers are forgiven (0.05%). In other words, we can replace the split LSF vectors of any given test sentence with split LSF vectors in the training set of approximately 5 sentences in such a manner that the average spectral distortion (SD) is less than 1.0 dB, there are (practically) no outliers having SD above 4 dB, and less than 2% of frames have SD between 2 and 4 dB.

Table 2. Spectral distortion using 5, 10, 20, 50, and 100 training sentences without splitting (*N*) and with 3-3-4 splitting (*S*).

	Trans-parent		5	10	20	50	100
Mean SD (dB)	<1.00	N	2.23	2.00	1.80	1.59	1.46
		S	0.80	0.63	0.51	0.38	0.31
2 dB (%) outliers	<2.00	N	58.5	46.1	34.7	21.4	14.0
		S	1.57	0.47	0.15	0.03	0.01
4 dB (%) outliers	0	N	2.63	0.95	0.38	0.10	0.04
		S	0.05	0.00	0	0	0

The above results suggest that it might be beneficial to use split LSF vectors in voice conversion if the aim is to be able to cope with small training sets. However, the design of the splitting scheme will have to be a trade-off between efficient space occupation and proper handling of correlations between the LSFs: the use of scalars would offer the most efficient space occupation but it would neglect the relationships between LSFs.

3. PROPOSED METHOD

Kain [4] proposed to combine the source vector x and the target vector y as $z = [x^T y^T]^T$ to estimate the GMM parameters (prior probability α_i, mean vector μ_i , and covariance matrix Σ_i) for each mixture $i = 1 \dots Q$. In conversion, the mapped target $\hat{y}$ is formed from the source x as

$$\hat{y} = \sum_{i=1}^{Q} h_i(x)[\mu_i^y + \Sigma_i^{yx}\Sigma_i^{xx-1}(x - \mu_i^x)] \quad (1)$$

where

$$\Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix} \text{ and } \mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix}$$

and

$$h_i(x) = \frac{\alpha_i \mathcal{N}(x; \mu_i^x, \Sigma_i^{xx})}{\sum_{j=1}^{Q} \alpha_j \mathcal{N}(x; \mu_j^x, \Sigma_j^{xx})}$$

Although Eq. 1 does not require that the dimension of the source feature is equal to that of target's, many voice conversion studies model jointly full vectors having equal dimension. We refer to this approach as full-vector GMM. Stylianou et al. [3] reported that using diagonal covariance matrices instead of full covariance matrices did not affect the results significantly. However, they were using MFCCs which do not exhibit strong correlation properties like LSFs. In addition, the k^{th} MFCC coefficient of the source corresponds to the k^{th} MFCC coefficient of the target but there may not always be such correspondence between the k^{th} source LSF and the k^{th} target LSF. Consequently, we have considered necessary to use full covariance matrices in LSF conversion.

In general, GMMs can model very complicated dependencies between variables if we assume that the size of the training data and the number of mixtures are not limited. Nevertheless, with a small amount of data, the conversion result may be really bad due to overfitting and due to the fact that it may not be possible to get reliable estimates for all the elements in the covariance matrix. Our idea is to utilize the inherent intra-frame correlation properties of LSFs. The first step in the process is to calculate the correlation coefficients between all the elements of the joint source-target LSF vectors. In the case of 10-dimensional LSFs, this results in a symmetric 20x20 matrix, as shown for one speaker pair in Table 1.

After computing the correlation coefficients based on the training data, it is checked for each target LSF which correlations are meaningful, i.e. the absolute value is above some threshold, say 0.5. For example for the fourth target LSF in Table 1 (corresponding to 14th element in the matrix), the most influential features would be source LSFs 2 and 3 and target LSFs 2, 3, and 5. Similarly, for the seventh target LSF we would choose source LSFs 5 and 6 and target LSFs 5, 6 and 8. A separate GMM is then built for each LSF using the feature elements having high correlation based on the source-target correlation coefficients. For each LSF model, it is ensured that at least one source LSF is included in the GMM estimation since the conversion phase would be impossible otherwise. Note that for example with the model of the first target LSF, the second target LSF could also be predicted but this option is not currently used in the conversion phase: for the second LSF, there is another GMM constructed based on its correlations.

The conversion is done using Eq. 1 that also holds for different combinations of data elements. For example, the training data for the fourth LSF would contain vectors of type $z = [x_2 x_3 y_2 y_3 y_4 y_5]^T$. In conversion, however, only the fourth LSF needs to be converted: only the third value in the mean vectors of μ_i^y of the model and the third row of matrix Σ_i^{xy} are needed in the conversion although the target features y_2, y_3, y_4 and y_5 are also used in the GMM training.

It is clear the proposed method cannot model complicated dependencies in the data but neither can the full-vector GMM when there is only a small amount of training data available. It is also known that the first LSFs are perceptually more important than the last ones (like ninth and tenth) but they are still treated equally in full-vector training. In the proposed approach, the different LSF elements can be treated differently e.g. by adjusting the correlation thresholds and the number of mixtures for the different GMMs. In general, there should be a perceptual error for training the LSF mapping function.

4. EVALUATION

For evaluation, we used a freely available CMU Arctic database [8] Mappings were trained for two pairs: *slt-bdl* (female to male) and *bdl-slt* (male to female). In analysis and synthesis, we used a voice conversion framework similar to the one presented in [11]. All the tests focused only on the differences in LSF conversion, and the conversion of all other parameters (voicing information and harmonic amplitudes for the residual spectrum, pitch and energy) were handled in an identical way in all tests. Only pitch level adjustment and residual spectrum resampling was carried out.

In the training phase, parallel utterances from the source and the target speakers were aligned using standard DTW based techniques. Silent frames and frame pairs with highly different levels of voicing were discarded. Moreover, one source frame was allowed to correspond to only one target frame in such a way that repeated source frames were combined by averaging their corresponding target LSFs. The GMMs, for both the proposed approach and the conventional approach, were trained using the EM algorithm with relative convergence rate threshold of $1e^{-6}$.

4.1. Objective results

While we acknowledge that objective measures are not always very reliable in voice conversion research, we did some objective measurements to get some numerical evidence on the usefulness of the proposed idea. As the study involved the conversion of the LSFs, all the measurements were based on LSFs. Because the use of root mean squared error of LSF vectors can be misleading, we used normalized cepstral distance (NCD) [6] and spectral distortion (SD) instead. NCD was calculated for 13th order MFCCs exluding the first MFCC as follows:

$$e(\hat{c}^t, c^t) = \frac{\sum_{i=1}^{N} \sum_{j=2}^{13} (\hat{c}_{ij}^t - c_{ij}^t)^2}{\sum_{i=1}^{N} \sum_{j=2}^{13} (c_{ij}^s - c_{ij}^t)^2} \quad (2)$$

where $\hat{c}^t$ is the predicted target, c^t is the real target vector, c^s is the source vector and N is the number of samples. We calculated the MFCCs from the LP spectrum based on the source, target and converted LSFs. SD was calculated for the band from 125 Hz to 3100 Hz.

As learnt from [6], the number of mixtures should be low when there is not much training data available and thus the full-vector GMM only had 4 and 8 mixtures whereas the split vector GMM had 2, 4, and 8 mixtures. The NCD and SD values given in Table 3 were computed using 50-sentence testing data sets not included in the training. The training data sets consisted of 2, 3, 5, 10 or 20 sentences and the results were averaged from 20 trials.

4.2. Listening test results

A preference test measuring the LSF conversion performance using training sets of only two parallel sentences was carried out. In the test, a full-vector GMM with 4 mixtures was compared against a split vector GMM with 2 mixtures. As discussed in the beginning of the section, the comparison only involved LSF conversion and all the other parameters were processed in identical way.

11 listeners participated in the test and and they were asked to give preference ratings for speech samples from the viewpoint of speech quality and speaker identity (which sample sounded more like the speaker in the reference target samples). The listeners could also answer "equal". The listening test included the same speaker

Table 3. Normalized cepstral distances (left of |) and mean spectral distortion in dB (right of |, in dB) for the training sets of 2, 5, 10 and 20 sentences, computed using the conventional method with 4 and 8 mixtures (*c-4* and *c-8*), and the proposed approach with 2, 4, and 8 mixtures (*sp-2*, *sp-4*, *sp-8*).

slt-bdl	2	5	10	20
c-4	0.65 | 5.1	0.51 | 4.6	0.47 | 4.3	0.44 | 4.2
c-8	0.80 | 5.5	0.55 | 4.7	0.48 | 4.4	0.44 | 4.2
sp-2	0.54 | 4.7	0.50 | 4.5	0.48 | 4.4	0.48 | 4.5
sp-4	0.55 | 4.8	0.49 | 4.5	0.47 | 4.4	0.47 | 4.4
sp-8	0.58 | 4.9	0.50 | 4.6	0.47 | 4.4	0.46 | 4.4
bdl-slt	2	5	10	20
c-4	0.55 | 4.7	0.42 | 4.2	0.39 | 3.9	0.37 | 3.8
c-8	0.75 | 5.2	0.46 | 4.3	0.40 | 4.0	0.37 | 3.8
sp-2	0.47 | 4.4	0.43 | 4.2	0.42 | 4.1	0.40 | 4.1
sp-4	0.47 | 4.5	0.42 | 4.2	0.41 | 4.1	0.41 | 4.1
sp-8	0.49 | 4.6	0.42 | 4.2	0.41 | 4.1	0.41 | 4.0

Table 4. Preference percentages for *quality* from a listening test.

Quality	Conventional	Proposed	Equal
slt-bdl	1.7%	78.4%	19.9%
bdl-slt	10.8%	69.9%	19.3%

pairs, *slt-bdl* and *bdl-slt*, as the objective tests. For both pairs, four different random two-sentence training sets were used for training the GMMs. The resulting GMMs were used to convert 4 different randomly-selected sentences, resulting in 32 pairs to evaluate. In total, the number of answers in the listening test was 352 for both quality and identity.

The results for the quality preference are shown in Table 4 and the closeness to the target identity in Table 5. The quality was found clearly better with the proposed method. As we expected, the identity related results show less difference since both methods are based on the use of GMMs that has restricted identity conversion capabilies due to the over-smoothing phenomenon.

5. DISCUSSION

As shown in Section 4, the proposed approach offers clear performance advantages when the amount of training data is very limited. This is not the only benefit. The method is also very flexible and it allows making reductions in the memory requirements and in the computational complexity. The model size can be adjusted by changing the correlation threshold or the number of mixtures in the GMMs. Moreover, different LSFs can be treated differently based on their perceptual relevance: for example, bigger models and/or more mixtures can be used for the more important first LSFs while smaller models can be used for the less important last LSFs. Considering the results in Table 3, is is likely that the most optimal split vector case was not included. Probably for the best result, different LSFs should have been modeled with different amount of mixtures. In addition, the correlation threshold can be different for different LSFs and it is also possible to model less important elements (e.g. the ninth and the tenth LSF) jointly. It should also be noted that lower mixture numbers can be used with the proposed approach than with the conventional full-vector GMMs because the dimension is smaller.

Even though the objective measurements suggest that the performance advantage of the proposed approach is lost with increasing training set size, it has been verified that perceptually the performance is still very close to that of the convetional full-vector GMM even with large training sets in the order of 100 sentences.

Table 5. Preference percentages for *identity* from a listening test.

Identity	Conventional	Proposed	Equal
slt-bdl	4.6%	38.6%	56.8%
bdl-slt	11.9%	31.8%	56.3%

6. CONCLUSIONS

We have presented a novel approach for LSF conversion that can cope with sparse training data. The proposed approach takes into account the inherent intra-frame correlation properties of LSFs. The objective measurements and the listening test results show the usefulness of the proposed approach, and demonstrate the benefits over the conventional joint density estimation based on full LSF vectors. The proposed method is especially useful if only a very small amount of training data is availble.

7. REFERENCES

[1] M. Abe, S. Nakamura, K.Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *ICASSP*, 1988, pp. 565–568.

[2] O. Turk and L.M. Arslan, "Robust processing techniques for voice conversion," *Computer Speech and Language*, vol. 4(20), pp. 441–467, October 2006.

[3] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Speech and Audio Processing*, vol. 6(2), pp. 131–142, March 1998.

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

[5] J. Yamagishi, K. Ogata, Y. Nakano, J. Isogai, and T. Kobayashi, "HSMM-based model adaptation algorithms for average-voice-based speech synthesis," in *ICASSP*, 2006, vol. I, pp. 77–80.

[6] L. Mesbahi, V. Barreaud, and O. Boeffard, "GMM-based speech transformation systems under data reduction," in *6th ISCA Speech Synthesis Workshop (SSW6)*, 2007, pp. 119–124.

[7] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," in *ICASSP*, 2007, vol. 4, pp. 1249–1252.

[8] J. Kominek and A.W. Black, "CMU Arctic databases for speech synthesis," Tech. Rep., Carnegie Mellon University, 2003.

[9] K. Paliwal and B. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. on Speech and Audio Processing*, vol. 1(1), pp. 3–14, January 1993.

[10] E. Helander, J. Nurminen, and M. Gabbouj, "Analysis of LSF frame selection for voice conversion," in *International conference on Speech and Computer*, 2007, pp. 651–656.

[11] J. Nurminen, V. Popa, J. Tian, Y. Tang, and I. Kiss, "A parametric approach for voice conversion," in *TC-STAR Workshop on Speech-to-Speech Translation*, 2006, pp. 225–229.