Super-Dirichlet Mixture Models using Differential Line Spectral Frequencies for Text-Independent Speaker Identification

Zhanyu Ma, Arne Leijon

KTH-Royal Institute of Technology
Sound and Image Processing Lab
Stockholm, Sweden
zhanyu@kth.se, leijon@kth.se

10.21437/Interspeech.2011-60

Abstract

A new text-independent speaker identification (SI) system is proposed. This system utilizes the line spectral frequencies (LSFs) as alternative feature set for capturing the speaker characteristics. The boundary and ordering properties of the LSFs are considered and the LSF are transformed to the differential LSF (DLSF) space. Since the dynamic information is useful for speaker recognition, we represent the dynamic information of the DLSFs by considering two neighbors of the current frame, one from the past frames and the other from the following frames. The current frame with the neighbor frames together are cascaded into a supervector. The statistical distribution of this supervector is modelled by the so-called super-Dirichlet mixture model, which is an extension from the Dirichlet mixture model. Compared to the conventional SI system, which is using the mel-frequency cepstral coefficients and based on the Gaussian mixture model, the proposed SI system shows a promising improvement.

Index Terms: Speaker recognition, differential line spectral frequencies, super-Dirichlet variable, mixture models

1. Introduction

Speaker recognition is an intensively studied area in the past decades [1, 2, 3]. Generally speaking, speaker recognition includes two tasks: 1) speaker identification (SI), to identify a particular speaker [4]; and 2) speaker verification (SV), to verify a speaker's claimed identity [5, 6]. Usually, the SI system contains three challenging phases: 1) extract features to represent the speakers's characteristics; 2) train separate models for each speaker to describe the statistical properties of the selected features; and 3) make decision by comparing the input data with the obtained models.

In the feature extraction phase, the mel-frequency cepstral coefficients (MFCCs) and the line spectral frequencies (LSFs) are widely used. The MFCCs are extracted from a short term spectrum of a windowed speech segment, based on a linear cosine transform of a log power spectrum modified non-linearly on the frequency domain [3, 7]. This feature suppresses the spectrum details in the high frequency area by introducing mel-scale filter bank, which takes the advantages of human ear's frequency selectivity. By considering the dynamic information contained over the time sequence, the velocity $\triangle$MFCCs and the acceleration $\triangle\triangle$MFCCs are always combined with the MFCCs to make a full dynamic feature representation. The LSFs were commonly utilized for quantizing the envelope of the linear predictive (LP) filter, because of its high efficiency of representing the LP coefficients [8]. As these filter parameters are determined by the speaker's articulatory system, they are used to convey the information about the speaker's identity. In some literature [3, 9], speaker recognition based on the LSF features were studied. Campbell [3] utilized the LSFs as features to make a speaker recognition system and showed that the LSF features are effective in the divergence shape measure. Cordeiro et al. [9] compared the LSFs, mel-LSFs, and MFCCs, with dynamic information, using support vector machines. From the human perception point of view, suppressing the high frequency information could lead to a better perceptual performance. However, if the speaker identification task is carried out based on the machine learned system, the information in the high frequency area might also contain some information which could help to recognize the speaker. In other words, the LSFs contains "full band" information and might perform better than the band-modified MFCCs. Thus, we utilize the LSFs as the features for the purposes of text-independent speaker identification. Furthermore, we transform the LSFs to the differential line spectral frequencies (DLSFs) to exploit the boundary and ordering properties of the LSFs [10]. To utilize the dynamic information over a time sequence, we combined the current DLSF feature frame with two neighbor frames, one from the past frames and the other one from the following frames, to build a supervector. Similar as the $\triangle$MFCCs and $\triangle\triangle$MFCCs, the neighbor frames contains dynamic information of the DLSFs over a time interval. In contrary to the $\triangle$MFCCs and $\triangle\triangle$MFCCs, which only contain the modified dynamic information, the neighbor DLSFs represent the "raw" dynamic information.

The Gaussian mixture model (GMM) is the most commonly used tool for modelling the statistical distribution of the data. In the context of speaker recognition [4, 6], the GMM and variate versions of GMM were widely used. In principle, the GMM could model arbitrary distributions, with unconstrained model complexity. However, for the data with bounded support or ordering property, some other distributions (*e.g.*, beta distribution, Dirichlet distribution) can model such type of data better and yield a better performance than the GMM based methods, (*e.g.*, image segmentation [11, 12], LSF modelling and quantization [10]). As we have obtained the supervectors to represent the dynamic information of the DLSF features, we model the distributions of the supervectors with a so-called super-Dirichlet mixture model (SDMM), which is extended from the Dirichlet mixture models. For each speaker, an SDMM was trained based on the training data. The identification decision was made by choosing the maximal log-likelihood of a test set againt all the trained speaker models.

We evaluate the proposed SI system with the TIMIT database [13], which contains sentences spoken by male and fe-

male speakers from different regions of the United States. The test sets were extracted with different durations (from 0.25s to 2s). Compared to the conventional GMM based SI system with MFCCs features, we will show that the proposed SDMM based SI system demonstrates a promising improvement.

2. Features for Speaker Identification

For the purpose of text-independent speaker identification, different features can be used to describe the vocal tract information of a speaker. The MFCCs and the LSFs are the most widely used features for speaker identification.

The MFCCs are extracted by the following procedure [7]:

1. take the Fourier transform of a windowed speech segment;

2. map the power spectrum onto the mel-scale, with *e.g.*, triangular overlapping window;

3. take the discrete cosine transform of the logarithm of the mel-scaled power spectrum.

For the speech segment (frame) at time t, we can extract K dimensional MFCC vector [1] as $\mathbf{c}(t) = [c_1(t), \ldots, c_K(t)]^T$. If we want to exploit the dynamic information of the MFCCs over time sequences, the velocity of the MFCCs can be approximated using linear regression as [7]

$$\triangle c_k(t) = \frac{\sum_{l=-\kappa}^{\kappa} l \cdot c_k(t+l)}{\sum_{l=-\kappa}^{\kappa} l^2},\ k = 1, \ldots, K. \qquad (1)$$

The acceleration $\triangle\triangle c_k(t)$ can be obtained in a similar way by replacing $c_k(t)$ in (1) by $\triangle c_k(t)$. A typical configuration is $\kappa = 3$ for $\triangle c_k(t)$ and $\kappa = 2$ for $\triangle\triangle c_k(t)$. Then the MFCCs dynamic feature supervector can be obtained as [7]

$$\mathbf{c}_{\text{sup}}(t) \triangleq \begin{bmatrix} \mathbf{c}(t) \\ \triangle\mathbf{c}(t) \\ \triangle\triangle\mathbf{c}(t) \end{bmatrix}. \qquad (2)$$

The LSFs are usually used for quantizing the LP filter coefficients [14]. In the linear predictive coding model, the filter $G(z)$ with order K is $G(z) = 1 + \sum_{k=1}^{K} a_k z^{-k}$. Then we can build two symmetric polynomials$P(z) = G(z) + z^{-(K+1)}G(z^{-1})$ and $Q(z) = G(z) - z^{-(K+1)}G(z^{-1})$. The zeros of $P(z)$ and $Q(z)$ are interleaved on the unit circle as $0 = \omega_{q_0} < \omega_{p_1} < \omega_{q_1} < \ldots < \omega_{q_{\frac{K}{2}}} < \omega_{p_{\frac{K}{2}+1}} = \pi$. Then the LSFs are defined as [14]

$$\mathbf{s} = [s_1, s_2, \ldots, s_K]^T = [\omega_{p_1}, \omega_{q_1}, \ldots, \omega_{p_{\frac{K}{2}}}, \omega_{q_{\frac{K}{2}}}]^T. \qquad (3)$$

Since s_k are strictly ordered and bounded in $[0, \pi]$, we can transform the LSF parameter $\mathbf{s}$ to DLSF $\mathbf{x}$ as [10]

$$\mathbf{x} = \varphi(\mathbf{s}) = A\mathbf{s}, \qquad (4)$$

where

$$A = \frac{1}{\pi} \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ -1 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & -1 & 1 \end{bmatrix}_{K \times K}.$$

By introducing $x_{K+1} = 1 - \sum_{k=1}^{K} x_k$, the DLSF $\mathbf{x} = [x_1, \ldots, x_{K+1}]^T$ [2] can be modelled by a Dirichlet distribution [10].

Similar as the dynamic information introduced to the MFCCs, we also consider the dynamic changes of the DLSFs over time. The velocity of the MFCCs calculated in (1) contains the information from 2κ adjacent frames. The operation of windowing and normalizing on the adjacent frames presents a modification of the dynamic information. In order to utilize a similar amount of dynamic information, we still consider the neighbor frames of the current frame $\mathbf{x}(t)$, but retain the "raw" information of the neighbor frames. In this paper, we only consider two neighbor frames, one from the past frames and one from the following frames. For instance, for the frame at time t, the neighbor frames are the frame extracted at time $t - \tau$ and $t + \tau$, which is

$$\mathbf{x}(t)_{\text{neighbor}} \triangleq \{\mathbf{x}(t-\tau), \mathbf{x}(t+\tau)\}, \qquad (5)$$

where τ is integer (*e.g.*, $\tau = 1$). We denote these frames as the τ-neighbor frames. With the τ-neighbor frames, a dynamic feature supervector can be obtained by arranging the current frame $\mathbf{x}(t)$ and the τ-neighbor frames as

$$\mathbf{x}_{\text{sup}}(t) \triangleq \begin{bmatrix} \mathbf{x}(t) \\ \mathbf{x}(t-\tau) \\ \mathbf{x}(t+\tau) \end{bmatrix} = \begin{bmatrix} x_{1,1}(t) \\ \vdots \\ x_{1,K+1}(t) \\ x_{2,1}(t) \\ \vdots \\ x_{2,K+1}(t) \\ x_{3,1}(t) \\ \vdots \\ x_{3,K+1}(t) \end{bmatrix}. \qquad (6)$$

We denote the element in the supervector $\mathbf{x}_{\text{sup}}(t)$ with three indices, $x_{n,k}(t)$ means this element is the kth element in the nth subvector (*i.e.*, $\mathbf{x}(t)$, $\mathbf{x}(t-\tau)$, or $\mathbf{x}(t+\tau)$) and t is the time index for the supervector $\mathbf{x}_{\text{sup}}(t)$. Thus the DLSF feature supervector has $3(K+1)$ dimensions. However, since each subvector has a redundant element, the supervector $\mathbf{x}_{\text{sup}}$ has the same degrees of freedom as $\mathbf{c}_{\text{sup}}$. The difference is $\mathbf{x}_{\text{sup}}$ contains "raw" information from the neighbor frames while $\mathbf{c}_{\text{sup}}$ modifies the information from the adjacent frames.

In principle, if both $\mathbf{x}_{\text{sup}}$ and $\mathbf{c}_{\text{sup}}$ considered the same dynamic range (*i.e.*, $\tau = 1$, and $\kappa = 1$ for both the velocity and the acceleration), these two feature representations should contain the same amount of information and contribute equally to the identification system. However, as the MFCCs suppressed the high frequency information according to the advantage of human's ear, it might throw away some high frequency information, which might be useful for machine to recognize a person but not helpful for human beings' perception. In this case, we conjecture that the DLSF feature supervector could perform better than the MFCCs supervector, or at least they should have the same performance.

Actually, the configuration of κ is usually greater than 1 for both the velocity and the acceleration. Then $\mathbf{c}_{\text{sup}}$ still has the chance to perform better than $\mathbf{x}_{\text{sup}}$, as we defined that $\mathbf{x}_{\text{neighbor}}$ contains only two neighbor frames.

To generalize the choice of neighbor features so that a broader dynamic range could be involved, it is still possible to

[1] For the purpose of speaker recognition, we discarded the first coefficient $c_0(t)$ when extracting the MFCCs [7].

[2] The Dirichlet variable with $K+1$ dimensions has K degrees of freedom, since the $(K+1)$th element is redundant.

choose more than two neighbor features. The cost of this choice is to increase the dimensionality of the feature supervector. Another possible way to make the neighbor choice more flexible is to choose asymmetric neighbor features, *i.e.*, to choose different offsets for the past frames and the following frames. The choice of neighbor features is flexible, which depends on the purpose of application.

In the following paragraphs, we only consider two neighbor features with symmetric offset τ. We choose the MFCCs and the DLSFs for the application of speaker identification and compare the performances. In fact, the experimental results in section 4 show that $\mathbf{x}_{\text{sup}}$ performs better than $\mathbf{c}_{\text{sup}}$, even with a narrower time spans.

3. Super-Dirichlet Mixture Model

As we have transformed the LSF parameter to the DLSF representation by exploiting the boundary and ordering properties, the Dirichlet distribution and the corresponding Dirichlet mixture model (DMM) could be utilized to model the statistical distribution of the DLSFs [12]. The dynamic feature supervector $\mathbf{x}_{\text{sup}}$ is actually a cascade of three DLSFs, then for each subvector in $\mathbf{x}_{\text{sup}}$ (*i.e.*, $\mathbf{x}(t)$, $\mathbf{x}(t-\tau)$, and $\mathbf{x}(t+\tau)$), it is Dirichlet distributed. Moreover, with a sequential DLSFs $\mathbf{x}(1),\ldots,\mathbf{x}(t),\ldots,\mathbf{x}(T)$, we usually assume that the DLSFs are independently generated from a underlying statistical model. Thus the subvectors in $\mathbf{x}_{\text{sup}}$ are independent from each other. Finally, we have the matrix which contains T supervectors (see (6)) as $\mathbf{X} = [\mathbf{x}_{\text{sup}}(1),\ldots,\mathbf{x}_{\text{sup}}(T)]$. We introduce a so-called super-Dirichlet distribution to model the statistical distribution of $\mathbf{x}_{\text{sup}}$. The probability density function (PDF) of the super-Dirichlet distribution is defined as

$$\mathbf{SDir}(\mathbf{x}_{\text{sup}};\boldsymbol{\alpha}) = \prod_{n=1}^{N} \frac{\Gamma\left(\sum_{k=1}^{K_n+1}\alpha_{n,k}\right)}{\prod_{k=1}^{K_n+1}\Gamma\left(\alpha_{n,k}\right)} \prod_{k=1}^{K_n+1} x_{n,k}^{\alpha_{n,k}-1},$$

where $\Gamma(\cdot)$ is the gamma function, N is the number of subvectors in the supervector, and K_n is the degrees of freedom of the nth subvector. $\alpha_{n,k}$ is the parameter corresponds to $x_{n,k}$. The PDF of the super-Dirichlet distribution is actually a multiplication of several PDFs of the Dirichlet distribution.

To model the multimodality of the distribution, we can apply the mixture model technique [15] to the super-Dirichlet distribution and obtain a super-Dirichlet mixture model (SDMM) to capture the DLSFs' underlying distribution as

$$f(\mathbf{X}) = \prod_{t=1}^{T}\sum_{m=1}^{M}\pi_m \mathbf{SDir}(\mathbf{x}_{\text{sup}}(t);\boldsymbol{\alpha}(m)), \qquad (7)$$

where π_m is the weighting factor and $\boldsymbol{\alpha}(m)$ denotes the parameter vector for the mth mixture component, respectively.

By extending the expectation maximization (EM) algorithm [10, 11] for the DMM, we introduce the EM algorithm for the SDMM in a similar way. In order to find the optimal value of the parameters by the EM approach, we consider $\mathbf{x}_{\text{sup}}(t)$ as the *incomplete* component labelled data. Also, we assign latent variables $\mathbf{z}_t = [z_{t1},\ldots,z_{tM}]^T$ as the *indication vector* with only one element equal to 1 and 0 for the rest ($z_{tm}=1$ indicates that the observation $\mathbf{x}_{\text{sup}}(t)$ is generated from the mth component). Now we consider $\mathbf{x}_{\text{sup}}(t)$ and $\mathbf{z}_t$ as the *complete* data. In the E step, the expected value of z_{tm} is estimated as

$$\overline{z}_{tm} = \mathbf{E}[z_{tm}] = \frac{\pi_m \mathbf{SDir}(\mathbf{x}_{\text{sup}}(t);\boldsymbol{\alpha}(m))}{\sum_{m=1}^{M}\pi_m \mathbf{SDir}(\mathbf{x}_{\text{sup}}(t);\boldsymbol{\alpha}(m))}. \qquad (8)$$

In the M step, we first calculate the weighting factor as $\pi_m = \frac{1}{T}\sum_{t=1}^{T}\overline{z}_{tm}$. For the mth mixture component, the parameter vector $\boldsymbol{\alpha}_m$ is divided into N subvectors and each parameter subvector corresponds to one subvector in $\mathbf{x}_{\text{sup}}$. Thus we estimate the nth parameter subvector as

$$\begin{aligned}&\begin{bmatrix} \psi(\alpha_{n,1}(m)) - \psi(\sum_{k=1}^{K_n+1}\alpha_{n,k}(m)) \\ \vdots \\ \psi(\alpha_{n,K_n+1}(m)) - \psi(\sum_{k=1}^{K_n+1}\alpha_{n,k}(m)) \end{bmatrix} \\ =&\frac{1}{\sum_{t=1}^{T}\overline{z}_{tm}}\begin{bmatrix} \sum_{t=1}^{T}\overline{z}_{tm}\log x_{n,1}(t) \\ \vdots \\ \sum_{t=1}^{T}\overline{z}_{tm}\log x_{n,K_n+1}(t) \end{bmatrix}\end{aligned} \qquad (9)$$

and this estimation should be done for all the N parameter subvectors. $\alpha_{n,k}(m)$ is defined in a similar way as $x_{n,k}(t)$. Then by cascading the subvectors together, the parameter supervector $\boldsymbol{\alpha}(m)$ can be obtained. Equation 9 can not be solved analytically. We apply the Newton-Raphson method to solve it numerically. For the DLSF feature supervector used in this paper, we have $N=3$ and $K_n = K$.

For the MFCCs, we model the $3K$ dimensional feature supervector with a GMM. For the convenience of implementation, we force the covariance matrix of each Gaussian component to be diagonal. Then for a GMM with M mixture components, the number of parameters is $M(6K+1)$ in total. For the SDMM with M mixture components, the number of parameters is only $M(3K+4)$. Thus the model complexity of the SDMM is smaller than GMM, given the same number of mixture components. Actually, we saved about half of the number of parameters when choosing the SDMM.

4. Experimental Results and Discussion

To verify the proposed SI system, we evaluate the speaker identification performance based on the TIMIT [13] speech database. The TIMIT database contains 630 male and female speakers and each speaker spoke ten sentences. During each round of evaluation, we randomly selected 25 speakers from the database.

In the training phase, seven sentences were randomly selected from one speaker as the training speech data. The speech in the training data set was segmented into frames with 25 ms duration and 10 ms stepsize. The silent frames were removed. For each frame, a hanning window was applied. As the speech data was sampled at the frequency of 16k Hz, we extracted $K=16$ order LP coefficients for each frame. Then the LP coefficients were transformed to the LSFs, and finally transformed to the DLSFs by (4). The dynamic 51 dimensional feature supervectors (with freedom of 48) for the DLSFs were obtained by the methods described in section 2. The obtained DLSFs feature supervectors were used to train a SDMM with M_{SD} mixture components for this specified speaker. Thus we have 25 SDMMs in total, one for each selected speaker.

In the test phase, the remaining three sentences from one speaker were processed in a similar way described above. A set of DLSFs feature supervectors was also obtained for each selected speaker. We randomly selected T consecutive frames from the test feature set. These selected frames were used as the feature set for identifying a speaker. Then 25 test feature sets, each contained T frames, were obtained for 25 speakers.

With the above described training-test procedures, we calculated 25 log-likelihoods of a test feature set from an anonymous speaker (which is one of the selected 25 speakers) against

Table 1: Comparison of identification rate (in %). $\tau = 1$.

Duration	0.25s	0.5s	1s	1.5s	2s
32-GMM & MFCCs	77.5	89.1	96.3	98.1	98.7
32-SDMM & DLSFs	79.5	89.6	97.3	98.7	99.3
48-SDMM & DLSFs	81.8	92.1	97.6	98.9	99.5

all the 25 trained speaker SDMM models. The trained model yielding the largest log-likelihood value was considered to have the same statistical property as the test feature set. Thus we assigned this anonymous speaker with the identity of this trained model. In each evaluation round, we randomly selected consecutive T frames from each speaker 10 times for identification. Thus we have $25 \times 10 = 250$ test sets to identify. The identification score of one evaluation round was calculated as the number of corrected identified test sets divided by the total number of test sets, which is $\frac{N_c}{250} \times 100\%$, where N_c means the number of correctly identified test sets. We ran evaluations of such rounds 30 times and the mean value is reported. Since the duration of test speech has effect on the identification result, we choose different speech durations [3] as $\{0.25, 0.5, 1, 1.5, 2\}$ seconds, and evaluated the performance of the proposed system respectively with the procedure above. Furthermore, we set the dynamic range $\tau = 1, 2$ respectively to demonstrate the effect of different time spans.

For a fair comparison, we also extracted $K = 16$ dimensional MFCCs with the methods described in section 2. Hanning window was used. With a similar procedure as the DLSFs, we modelled the 48 dimensional dynamic MFCCs feature supervectors from one speaker by a GMM with M_G mixture components. The identification scores were also evaluated by randomly selecting 25 speakers, with 30 rounds of evaluations.

Each mixture component in the GMM has $6K$ parameters, and each mixture component in the SDMM has $3(K + 1)$ parameters. We compared the two SI systems with the same number of mixture components, which is $M_G = M_{SD} = 32$. In this case, when $K = 16$, each GMM has $32 \times (6 \times 16 + 1) = 3104$ parameters, while each SDMM has $32 \times (3 \times (16 + 1) + 1) = 1664$ parameters. The model complexity of SDMM is about 54% of GMM. The comparison results are shown in Tab. 1. It is observed that the SDMM performs slightly better than the GMM, at different speech durations. However, the improvement is small. Therefore, we increased M_{SD} from 32 to 48, which gave $48 \times (3 \times (16 + 1) + 1) = 2496$ parameters to the SDMM. In this case, the model complexity of SDMM is about 80% of GMM, which is still significantly smaller. As shown in the last row of Tab. 1, a promising improvement was obtained.

To investigate the effect of different time spans, the comparison results with $\tau = 2$ are shown in Tab. 2. Even though an improvement was achieved, this improvement is less than the case with $\tau = 1$. Thus we conclude that the closer neighbor frames contribute more to the dynamic information.

In the above comparisons, the proposed SI system showed a promising improvement over the conventional one. We conjecture there are two main reasons: one reason is to model the DLSF coefficients with the SDMM, which exploited the boundary and ordering properties, even with a lower model complexity. The other reason is the DLSF representation, which contains the "full band" information, where as the MFCCs represent a smoothed version of the spectrum. Even though the human beings are not sensitive to the high frequency information, the machine could utilize the information in the high frequency area for recognition. The identification results might be further

[3] For a test speech with T frames, the duration is about $0.01T$, since the segment step size is 0.01s.

Table 2: Comparison of identification rate (in %). $\tau = 2$.

Duration	0.25s	0.5s	1s	1.5s	2s
32-GMM & MFCCs	77.5	89.1	96.3	98.1	98.7
32-SDMM & DLSFs	78.9	89.3	96.4	98.5	99.0
48-SDMM & DLSFs	78.2	90.0	96.6	98.7	99.3

improved by using additional features, but that is beyond the scope of this paper.

5. Conclusion

A new speaker identification system was proposed. This system is based on the DLSF representation of the speaker's vocal tract. The DLSF supervectors, which contain both the static and dynamic information of the DLSF coefficients, are used as features to identify a speaker. To exploit the boundary and ordering properties of the DLSF coefficients, the Dirichlet mixture model was extended to the so-called super Dirichlet mixture model, which is capable of modelling the statistical distribution of the DLSF supervector efficiently. Compared to the conventional MFCCs and GMM based SI system, the proposed SI system performs better at a wide range of test sample durations and with a lower model complexity.

6. References

[1] G. Doddington, "Speaker recognition: Identifying people by their voices," *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1651 – 1664, 1985.

[2] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: a feature-based approach," *Signal Processing Magazine, IEEE*, vol. 13, no. 5, p. 58, Sep. 1996.

[3] J. Campbell, J.P., "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437 –1462, Sep. 1997.

[4] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72 –83, Jan. 1995.

[5] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Process.*, vol. 2004, pp. 430–451, January 2004.

[6] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308 – 311, May 2006.

[7] J. Benesty and Y. Sondhi, M. M.and Huang, Eds., *Springer Handbook on Speech Processing*. Springer, 2008.

[8] K. K. Paliwal and W. B. Kleijn, *Speech Coding and Synthesis*. Amsterdam: Elsevier, 1995, ch. Quantization of LPC parameters, pp. 433–466.

[9] H. Cordeiro and C. Ribeiro, "Speaker characterization with MLSFs," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, 2006, pp. 1–4.

[10] Z. Ma and A. Leijon, "Modeling speech line spectral frequencies with Dirichlet mixture models," in *Proceedings of Interspeech*, 2010.

[11] N. Bouguila, D. Ziou, and J. Vaillancourt, "Unsupervised learning of a finite mixture model based on the Dirichlet distribution and its application," *Image Processing, IEEE Transactions on*, vol. 13, no. 11, pp. 1533–1543, Nov. 2004.

[12] Z. Ma and A. Leijon, "Bayesian estimation of beta mixture models with variational inference," *Pattern Analysis and Machine Intelligence, IEEE Transactions on. Accepted and to appear.*, 2011.

[13] "DARPA-TIMIT," *Acoustic-phonetic continuous speech corpus, NIST Speech Disc 1.1-1*, 1990.

[14] F. Soong and B. Juang, "Line spectrum pair (LSP) and speech data compression," in *Acoustics, Speech, and Signal Processing, 1984. IEEE International Conference on.*, vol. 9, Mar 1984, pp. 37–40.

[15] G. J. McLachlan and D. Peel, *Finite Mixture Models* Wiley, 2000.