Published in IET Signal Processing
Received on 5th July 2013
Revised on 4th March 2014
Accepted on 21st March 2014
doi: 10.1049/iet-spr.2013.0270

ISSN 1751-9675

Skew Gaussian mixture models for speaker recognition

Avi Matza, Yuval Bistritz
School of Electrical Engineering, Tel-Aviv University
E-mail: avimatza@post.tau.ac.il

Abstract: Gaussian mixture models (GMMs) are widely used in speech and speaker recognition. This study explores the idea that a mixture of skew Gaussians might capture better feature vectors that tend to have skew empirical distributions. It begins with deriving an expectation maximisation (EM) algorithm to train a mixture of two-piece skew Gaussians that turns out to be not much more complicated than the usual EM algorithm used to train symmetric GMMs. Next, the algorithm is used to compare skew and symmetric GMMs in some simple speaker recognition experiments that use Mel frequency cepstral coefficients (MFCC) and line spectral frequencies (LSF) as the feature vectors. MFCC are one of the most popular feature vectors in speech and speaker recognition applications. LSF were chosen because they exhibit significantly more skewed distribution than MFCC and because they are widely used [together with the related immittance spectral frequencies (ISF)] in speech transmission standards. In the reported experiments, models with skew Gaussians performed better than models with symmetric Gaussians and skew GMMs with LSF compared favourably with both skew symmetric and symmetric GMMs that used MFCC.

1 Introduction

Finite mixture of probability density functions (pdf) provide an effective mathematical tool for modelling a variety of statistical phenomena in diverse scientific fields such as astronomy, bio-genetics, medicine, economics and more [1]. The most widely used pdf in such mixtures is the Gaussian density function because of its simple structure and the fact that its parameters can be easily trained. Gaussian mixture models (GMMs) are widely used also in speaker recognition [2, 3] and speech recognition (as part of the hidden Markov model paradigm).

In this paper we explore the idea of replacing the Gaussian pdf in these mixture models with skew Gaussian pdfs. Our initiative stems from the following observations. Most GMM based speaker recognition systems use the so called mel frequency cepstral coefficients (MFCC) and its derivatives as feature vectors. These vectors have relatively symmetric empirical distributions, as illustrated in Fig. 1, and thus are well suited for modelling with symmetric Gaussians. However, the same cannot be said for the histograms in Fig. 2, which presents empirical distributions of the so called line spectral frequencies (LSF). LSF are features used to encode the spectral envelope of speech in many speech coder standards, see for example, [4, 5]. Some more recent wide band coders, for example, [6], use the related immittance spectral frequencies (ISF) [7] that feature similarly skewed distributions. The LSF and ISF are known to be strongly correlated to the formant frequencies of voiced speech. This correlation and their significant role in compressing speech, inspired several attempts to use them also for speaker recognition applications [8–10], but they were unsuccessful in providing the good discrimination capabilities that might have been expected. Consequently, this paper examines to what extent the success of GMM with MFCC is related to the relative non-skewed shape of MFCC, and whether the LSF can perform better with skew GMMs.

The paper first derives an expectation-maximisation (EM) algorithm to train a two-piece skew GMM. Then it presents the results of some basic speaker recognition experiments that compare the performance of symmetric and skew GMMs using MFCC and LSF as feature vectors. These experiments were held with relatively low order models that used short feature vectors and were trained and tested using limited amount of data. In these experiments, models with skew Gaussians performed, in general, better than symmetric Gaussians, and skew GMMs with LSF compared favourably with both skew and symmetric GMMs with MFCC. Some of the results in this paper were presented previously at a conference [11]. Specifically, the conference paper brings no proof for the used training algorithm and contains a smaller set of experiments (only clean speech with only the set of lower order models).

The paper is structured as follows, Section 2 describes the two-piece skew Gaussian distribution. Section 3 derives an expectation maximisation (EM) algorithm to train the parameters of a skew GMM distribution. Section 4 presents and analyses the results of the experiments and the paper concludes with some remarks.

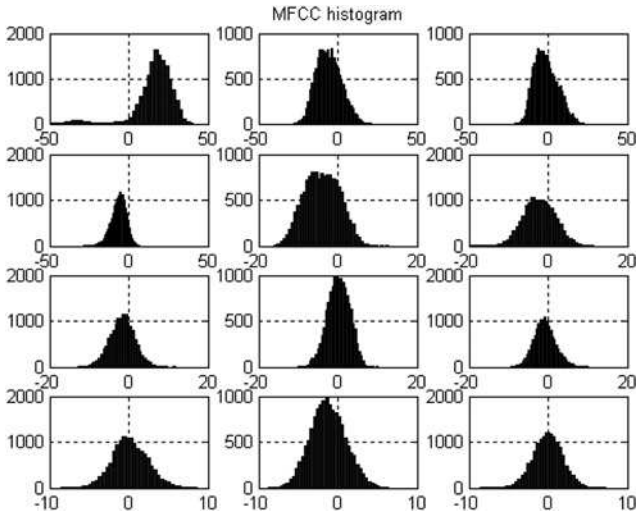

Fig. 1 *Histograms of 1 : 12 mell frequency cepstral coefficients*

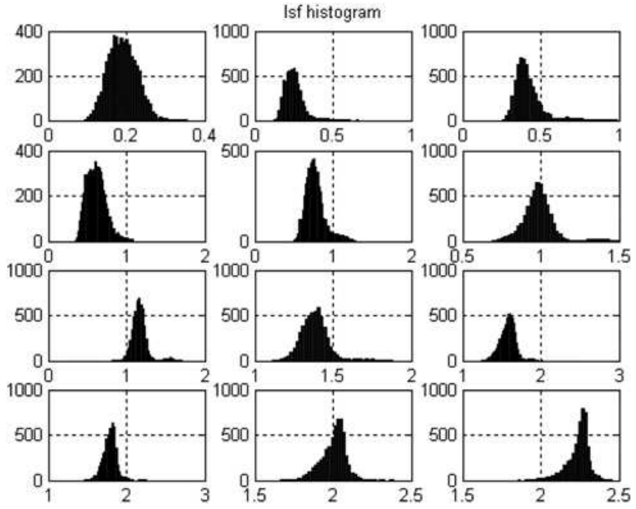

Fig. 2 *Histograms of 12 line frequency coefficients*

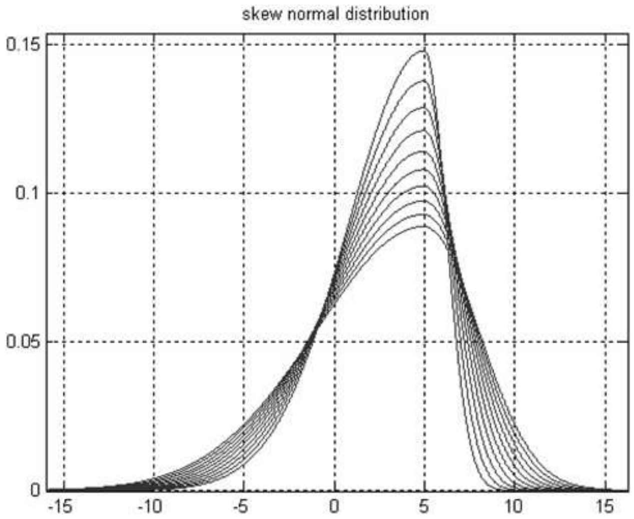

Fig. 3 *Two-piece skew Gaussian density functions with $\xi = 5$ and various $\acute{\sigma}$ and $\grave{\sigma}$*

2 The skew Gaussian distribution

The two-piece skew Gaussian pdf is defined as follows

$$a(x;\, \xi,\, \acute{\sigma},\, \grave{\sigma}) = \frac{2}{(\acute{\sigma} + \grave{\sigma})} \times \left(\phi\left[\frac{x-\xi}{\acute{\sigma}}\right] I(x \geq \xi) + \phi\left[\frac{x-\xi}{\grave{\sigma}}\right] I(x < \xi) \right) \tag{1}$$

where $\phi(x)$ denotes the standard Gaussian pdf, that is, $\phi(x) = (2\pi)^{-1/2}\mathrm{e}^{-x^2/2}$, and $I(c)$ is an indicator function that is equal to 1 if c is *true* and 0 if c is *false*. This representation has three-parameters: a location parameter ξ and two parameters, $\acute{\sigma}$ and $\grave{\sigma}$, that determine the shape of the pdfs tails to the right and left of the location parameter as illustrated in Fig. 3. When $\acute{\sigma} = \grave{\sigma} = \sigma$, the skew Gaussian pdf reduces to the standard symmetric Gaussian density function with mean ξ and covariance σ^2. The earliest appearance of this presentation is possibly in [12] where Gibbons and Mylroie suggested it to better fit certain asymmetrical experimental data. It was also introduced later by John [13]. This two-piece three-parameter presentation is not the only form in which the skew Gaussian appears in the literature. Another interesting representation, that was introduced by Azzalini [14, 15] and studied subsequently intensively by him and other researchers, uses three different parameters that express it as a product of $\phi(x)$ and its cumulative distribution function. For an account on these two ways of presenting skew Gaussian pdfs (as well as other skew distributions) see [16] and the references there in.

The pdf (1) is illustrated in Fig. 3. Note that when $\acute{\sigma} \neq \grave{\sigma}$ the location parameter ξ differs from the mean $E(x)$. The mean and the second and third central moments for $a(x;\, \xi,\, \acute{\sigma},\, \grave{\sigma})$ were reported in [12] and [13]. The mean is given by

$$E(x) = \xi + \sqrt{\frac{2}{\pi}}(\acute{\sigma} - \grave{\sigma}) \tag{2}$$

The variance is given by

$$\mathrm{Var}(x) = E\{[x - E(x)]^2\} = \left(1 - \frac{2}{\pi}\right)(\acute{\sigma} - \grave{\sigma})^2 + \acute{\sigma}\grave{\sigma} \tag{3}$$

and the third central moment is

$$E\{[x - E(x)]^3\} = \sqrt{\frac{2}{\pi}}(\acute{\sigma} - \grave{\sigma}) \times \left[\left(\frac{4}{\pi} - 1\right)(\acute{\sigma} - \grave{\sigma})^2 + \acute{\sigma}\grave{\sigma}\right] \tag{4}$$

The latter bears no specific name but it is closely related to a term called *skewness* which is a measure of the asymmetry of a probability distribution, see for example, [16], and defined such that zero skewness implies in this case that the pdf reduces to the symmetric Gaussian.

3 Skew Gaussian mixture models

Consider a skew GMM with K components for a D dimensional vector of random variables $\boldsymbol{x}_t = [x_{t1}, \ldots, x_{tD}]^t$

$$A(\boldsymbol{x}_t|\Lambda) = \sum_{k=1}^{K} w_k S_k(\boldsymbol{x}_t|\lambda_k) \tag{5}$$

where $S_k(\boldsymbol{x}_t|\lambda_k)$ is a multivariate skew Gaussian pdf that depends on some set of parameters λ_k, and $w_k > 0$ are the mixture weights such that $\sum_{k=1}^{K} w_k = 1$. The term Λ denotes the complete set of parameters that are used to define the model, namely $\Lambda = \{\lambda_k, w_k, k = 1, \ldots, K\}$.

Next we derive an EM algorithm for training the parameters of Λ. The derivation begins in a familiar manner [17] until the specific skew GMM parameters have to be introduced. Let $\mathcal{X} = \{\boldsymbol{x}_t, t = 1, \ldots, T\}$ be the observed data (feature vectors of length D) that will be regarded as independent and identically distributed (iid) random variables. Then, the probability that the model assigns to it is

$$P(\mathcal{X}|\Lambda) = \prod_{t=1}^{T} A(\boldsymbol{x}_t|\Lambda)$$

where $S_k(\boldsymbol{x}_t|\lambda_k)$, in (5), is any arbitrary pdf (not necessarily the multivariate skew Gaussian) that depends on a set of parameters λ_k. The goal of the training procedure is to find the set of parameters Λ that maximises the probability $P(\mathcal{X}|\Lambda)$.

The derivation of the EM algorithm usually assumes that besides the observed vectors $\boldsymbol{x}_t \in \mathcal{X}$ there are some additional 'unobserved' data $y_t \in \mathcal{Y}$. The algorithm consists of two parts, the expectation E-step and a maximisation M-step. At the expectation step, the conditional expected value (with respect to the unobserved data) of the log-likelihood function for the 'complete' data set $(\mathcal{X}, \mathcal{Y})$ is formed given the current estimate Λ^o of the underlying parameters and the given data. Let us denote it by

$$Q(\Lambda, \Lambda^o) = E_y\left[\log P(\mathcal{X}, \mathcal{Y}|\Lambda)|\mathcal{X}, \Lambda^o\right] \tag{6}$$

The M-step that follows, maximises $Q(\Lambda, \Lambda^o)$ with respect to the estimated parameters Λ, as a means to find Λ^{opt}, which is the optimal (maximum likelihood) estimation of Λ

$$\Lambda^{\text{opt}} = \arg\max_{\Lambda} Q(\Lambda, \Lambda^o) \tag{7}$$

Let $f(\mathcal{Y}|\mathcal{X}, \Lambda^o)$ denote the joint marginal distribution for all y, the above $Q(\Lambda,\Lambda^o)$ can be spelled out as

$$Q(\Lambda, \Lambda^o) = \int_{y\in\mathcal{Y}} \log P(\mathcal{X}, \mathcal{Y}|\Lambda) f(\mathcal{Y}|\mathcal{X}, \Lambda^o)\,\mathrm{d}y \tag{8}$$

Using the assumption that the observed data $\boldsymbol{x}_t$ are iid and Bayes law, the term $\log P(\mathcal{X}, \mathcal{Y}|\Lambda)$ can be expressed by

$$\begin{aligned} \log P(\mathcal{X}, \mathcal{Y}|\Lambda) &= \log \prod_{t=1}^{T} P(\boldsymbol{x}_t, y_t|\Lambda) \\ &= \sum_{t=1}^{T} \log\left[P(\boldsymbol{x}_t|y_t, \Lambda)P(y_t|\Lambda)\right] \end{aligned} \tag{9}$$

Assume now that the pdf of each $\boldsymbol{x}_t$ is given by the mixture (5). Then, each 'unobserved' data $y_t \in \mathcal{Y}$ can be chosen as an index $y_t \in [1, \ldots, K]$ that points to a specific component in the assumed mixture. Therefore the probability of y_t given all other parameters is equal to the probability of selecting the mixture component whose index is y_t, that is, $P(y_t|\Lambda) = w_{y_t}$. In this case (9) becomes

$$\begin{aligned} \log P(\mathcal{X}, \mathcal{Y}|\Lambda) &= \sum_{t=1}^{T} \log\left[P(\boldsymbol{x}_t|y_t, \Lambda)P(y_t|\Lambda)\right] \\ &= \sum_{t=1}^{T} \log\left[w_{y_t} S_{y_t}\left(\boldsymbol{x}_t|\lambda_{y_t}\right)\right] \end{aligned} \tag{10}$$

and $S_{y_t}\left(\boldsymbol{x}_t|\lambda_{y_t}\right)$ is the probability of $\boldsymbol{x}_t$ to come from the mixture component associated with y_t given the distribution's parameters. The marginal distribution $f(\mathcal{Y}|\mathcal{X}, \Lambda^o)$ in (8) can be written as

$$f(\mathcal{Y}|\mathcal{X}, \Lambda^o) = \prod_{j=1}^{T} P(y_j|\boldsymbol{x}_j, \Lambda^o) \tag{11}$$

Setting (10) and (11) into (8), and using the fact that y_t takes discrete values to replace the integral by a sum, $Q(\Lambda, \Lambda^o)$ becomes

$$Q(\Lambda, \Lambda^o) = \sum_{y\in\boldsymbol{y}} \left\{ \sum_{t=1}^{T} \log\left[w_{y_t} S_{y_t}\left(\boldsymbol{x}_t|\lambda_{y_t}\right)\right] \prod_{j=1}^{T} P(y_j|\boldsymbol{x}_j, \Lambda^o) \right\} \tag{12}$$

Expand the above expression as follows

$$\begin{aligned} Q(\Lambda, \Lambda^o) &= \sum_{y_1=1}^{K}\sum_{y_2=1}^{K}\cdots\sum_{y_{T-1}=1}^{K}\sum_{y_T=1}^{K} \\ &\quad \times \left\{ \sum_{t=1}^{T} \log\left[w_{y_t} S_{y_t}\left(\boldsymbol{x}_t|\lambda_{y_t}\right)\right] \prod_{j=1}^{T} P(y_j|\boldsymbol{x}_j, \Lambda^o) \right\} \\ &= \sum_{y_1=1}^{K}\sum_{y_2=1}^{K}\cdots\sum_{y_{T-1}=1}^{K}\sum_{y_T=1}^{K} \\ &\quad \times \left\{ \sum_{t=1}^{T}\sum_{k=1}^{K} \delta_{k,y_t} \log\left[w_k S_k(\boldsymbol{x}_t|\lambda_k)\right] \prod_{j=1}^{T} P(y_j|\boldsymbol{x}_j, \Lambda^o) \right\} \\ &= \sum_{t=1}^{T}\sum_{k=1}^{K} \log\left[w_k S_k(\boldsymbol{x}_t|\lambda_k)\right] \\ &\quad \times \sum_{y_1=1}^{K}\sum_{y_2=1}^{K}\cdots\sum_{y_{T-1}=1}^{K}\sum_{y_T=1}^{K} \delta_{k,y_t} \prod_{j=1}^{T} P(y_j|\boldsymbol{x}_j, \Lambda^o) \end{aligned} \tag{13}$$

Proceed with further simplification of the following term

$$
\begin{aligned}
&\sum_{y_1=1}^{K}\sum_{y_2=1}^{K}\cdots\sum_{y_{T-1}=1}^{K}\sum_{y_T=1}^{K}\delta_{k,y_t}\prod_{j=1}^{T}P(y_j|\boldsymbol{x}_j,\Lambda^o)\\
&=\left\{\sum_{y_1=1}^{K}\cdots\sum_{y_{t-1}=1}^{K}\sum_{y_{t+1}=1}^{K}\cdots\sum_{y_T=1}^{K}\prod_{j=1,j\neq t}^{T}P(y_j|\boldsymbol{x}_j,\Lambda^o)\right\}P(k|\boldsymbol{x}_t,\Lambda^o)\\
&=\prod_{j=1,j\neq t}^{T}\left[\sum_{y_j=1}^{K}P(y_j|\boldsymbol{x}_j,\Lambda^o)\right]P(k|\boldsymbol{x}_t,\Lambda^o)=P(k|\boldsymbol{x}_t,\Lambda^o)
\end{aligned}
\tag{14}
$$

since $\sum_{y_j=1}^{K}P(y_j|\boldsymbol{x}_j,\Lambda^o)=1$.

Inserting (14) into (13) we finally obtain

$$
\begin{aligned}
Q(\Lambda,\Lambda^o)&=\sum_{t=1}^{T}\sum_{k=1}^{K}\log\left[w_kS_k(\boldsymbol{x}_t|\lambda_k)\right]P(k|\boldsymbol{x}_t,\Lambda^o)\\
&=\sum_{t=1}^{T}\sum_{k=1}^{K}\log(w_k)P(k|\boldsymbol{x}_t,\Lambda^o)\\
&\quad+\sum_{t=1}^{T}\sum_{k=1}^{K}\log\left[S_k(\boldsymbol{x}_t|\lambda_k)\right]P(k|\boldsymbol{x}_t,\Lambda^o)
\end{aligned}
\tag{15}
$$

The maximisation of $Q(\Lambda,\Lambda^o)$ can be carried out by maximising separately the two parts of (15); the first part for w_k and the second part for λ_k. In order to maximise the first part, we introduce a Lagrange multiplier to guarantee the constraint $\sum_{k=1}^{K}w_k=1$ and seek a solution to

$$
\frac{\partial}{\partial w_k}\left[\sum_{k=1}^{K}\sum_{t=1}^{T}\log(w_k)P(k|\boldsymbol{x}_t,\Lambda^o)+\mu\left(-1+\sum_{k=1}^{K}w_k\right)\right]=0
\tag{16}
$$

for $k=1,\ldots,K$, that becomes

$$
\sum_{t=1}^{T}\frac{1}{w_k}P(k|\boldsymbol{x}_t,\Lambda^o)+\mu=0
\tag{17}
$$

Isolate w_k and summing over k, using $\sum_{k=1}^{K}w_k=1$, gives $\mu=-T$. Solving for w_k we obtain

$$
w_k=\frac{1}{T}\sum_{t=1}^{T}P(k|\boldsymbol{x}_t,\Lambda^o)
\tag{18}
$$

Note that the derivation so far did not involve any assumptions regarding the density function $S_k(\boldsymbol{x}_t|\lambda_k)$ in the mixture (5). However, the maximisation of the second part of (15) requires specification of $S_k(\boldsymbol{x}_t|\lambda_k)$. We assume that each entry x_{td} of an observation vector $\boldsymbol{x}_t=[x_{t1},\ldots,x_{tD}]^t$ has skew Gaussian pdf, $x_{td}\sim a(x;\xi,\acute{\sigma},\grave{\sigma})$ and that the entries are independent so that we can write

$$
S_k(\boldsymbol{x}_t|\lambda_k)=\prod_{d=1}^{D}a\left(x_{td};\xi_{kd},\acute{\sigma}_{kd},\grave{\sigma}_{kd}\right)
\tag{19}
$$

The second part (15) becomes after inserting (19) into it (see (20))

where

$$
P(k|\boldsymbol{x}_t,\Lambda^o)=\frac{w_k^o\prod_{d=1}^{D}a\left(x_{td};\xi_{kd}^o,\acute{\sigma}_{kd}^o,\grave{\sigma}_{kd}^o\right)}{\sum_{i=1}^{K}w_i^o\prod_{d=1}^{D}a\left(x_{td};\xi_{id}^o,\acute{\sigma}_{id}^o,\grave{\sigma}_{id}^o\right)}
\tag{21}
$$

and $C=\log2-(1/2)\log2\pi$ is a constant that does not affect the maximisation.

Using (20) to differentiate $Q(\Lambda,\Lambda^o)$ with respect to ξ_{kd} gives

$$
\begin{aligned}
&\sum_{t=1}^{T}\left[\frac{x_{td}-\xi_{kd}}{\acute{\sigma}_{kd}^2}I(x_{td}\geq\xi_{kd})+\frac{x_{td}-\xi_{kd}}{\grave{\sigma}_{kd}^2}I(x_{td}<\xi_{kd})\right]\\
&\quad\times P(k|\boldsymbol{x}_t,\Lambda^o)=0
\end{aligned}
\tag{22}
$$

The maximum is reached at ξ_{kd} for which the derivative vanishes (see (23))

Note that in each of the summations over $1\leq t\leq T$ that appear in the expression (23), the sum contains only terms for which x_{td} fulfils the requirement shown in the corresponding indicator function.

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{k=1}^{K}\log\left[S_k(\boldsymbol{x}_t|\lambda_k)\right]P(k|\boldsymbol{x}_t,\Lambda^o)\\
&=\sum_{t=1}^{T}\sum_{k=1}^{K}\log\left[\prod_{d=1}^{D}a\left(x_{td};\xi_{kd},\acute{\sigma}_{kd},\grave{\sigma}_{kd}\right)\right]P(k|\boldsymbol{x}_t,\Lambda^o)\\
&=\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{d=1}^{D}\left\{C-\log\left(\acute{\sigma}_{kd}+\grave{\sigma}_{kd}\right)-\frac{1}{2}\left[\frac{x_{td}-\xi_{kd}}{\acute{\sigma}_{kd}}\right]^2I(x_{td}\geq\xi_{kd})-\frac{1}{2}\left[\frac{x_{td}-\xi_{kd}}{\grave{\sigma}_{kd}}\right]^2I(x_{td}<\xi_{kd})\right\}P(k|\boldsymbol{x}_t,\Lambda^o)
\end{aligned}
\tag{20}
$$

$$
\xi_{kd}=\frac{\grave{\sigma}_{kd}^2\sum_{t=1}^{T}x_{td}P(k|\boldsymbol{x}_t,\Lambda^o)I(x_{td}\geq\xi_{kd})+\acute{\sigma}_{kd}^2\sum_{t=1}^{T}x_{td}P(k|\boldsymbol{x}_t,\Lambda^o)I(x_{td}<\xi_{kd})}{\grave{\sigma}_{kd}^2\sum_{t=1}^{T}P(k|\boldsymbol{x}_t,\Lambda^o)I(x_{td}\geq\xi_{kd})+\acute{\sigma}_{kd}^2\sum_{t=1}^{T}P(k|\boldsymbol{x}_t,\Lambda^o)I(x_{td}<\xi_{kd})}
\tag{23}
$$

Next, use (20) to differentiate $Q(\Lambda, \Lambda^o)$ with respect to $\acute{\sigma}_{kd}$

$$\sum_{t=1}^{T}\left[\frac{-1}{\acute{\sigma}_{kd}+\grave{\sigma}_{kd}}+\frac{(x_{td}-\xi_{kd})^2}{\grave{\sigma}_{kd}^3}I(x_{td}<\xi_{kd})\right]P(k|\boldsymbol{x}_t, \Lambda^o) \tag{24}$$

Setting the derivative to zero gives the equation

$$\begin{aligned}&\grave{\sigma}_{kd}^3\sum_{t=1}^{T}P(k|\boldsymbol{x}_t, \Lambda^o)-\grave{\sigma}_{kd}\sum_{t=1}^{T}(x_{td}-\xi_{kd})^2\\ &\quad\times P(k|\boldsymbol{x}_t, \Lambda^o)I(x_{td}<\xi_{kd})\\ &\quad-\acute{\sigma}_{kd}\sum_{t=1}^{T}(x_{td}-\xi_{kd})^2P(k|\boldsymbol{x}_t, \Lambda^o)I(x_{td}<\xi_{kd})=0\end{aligned} \tag{25}$$

Note that the term $\grave{\sigma}_{kd}^3\sum_{t=1}^{T}P(k|\boldsymbol{x}_t, \Lambda^o)$ is not multiplied by an indicator $I(c)$, namely, it is calculated over the whole sample space. Similarly, differentiating $Q(\Lambda, \Lambda^o)$ with respect to $\acute{\sigma}_{kd}$ gives a similar dual equation

$$\begin{aligned}&\acute{\sigma}_{kd}^3\sum_{t=1}^{T}P(k|\boldsymbol{x}_t, \Lambda^o)-\acute{\sigma}_{kd}\sum_{t=1}^{T}(x_{td}-\xi_{kd})^2\\ &\quad\times P(k|\boldsymbol{x}_t, \Lambda^o)I(x_{td}\geq\xi_{kd})\\ &\quad-\grave{\sigma}_{kd}\sum_{t=1}^{T}(x_{td}-\xi_{kd})^2P(k|\boldsymbol{x}_t, \Lambda^o)I(x_{td}\geq\xi_{kd})=0\end{aligned} \tag{26}$$

To simplify the representation, we introduce the following notation

$$\tilde{C}_k=\sum_{t=1}^{T}P(k|\boldsymbol{x}_t, \Lambda^o) \tag{27}$$

$$\overleftarrow{C}_{kd}=\sum_{t=1}^{T}(x_{td}-\xi_{kd})^2P(k|\boldsymbol{x}_t, \Lambda^o)I(x_{td}<\xi_{kd}) \tag{28}$$

$$\overrightarrow{C}_{kd}=\sum_{t=1}^{T}(x_{td}-\xi_{kd})^2P(k|\boldsymbol{x}_t, \Lambda^o)I(x_{td}\geq\xi_{kd}) \tag{29}$$

With them the pair of equations for $\grave{\sigma}_{kd}$ and $\acute{\sigma}_{kd}$ become

$$\begin{aligned}\tilde{C}_k\,\grave{\sigma}_{kd}^3-\overleftarrow{C}_{kd}\,\grave{\sigma}_{kd}-\overleftarrow{C}_{kd}\,\acute{\sigma}_{kd}=0\\ \tilde{C}_k\,\acute{\sigma}_{kd}^3-\overrightarrow{C}_{kd}\,\acute{\sigma}_{kd}-\overrightarrow{C}_{kd}\,\grave{\sigma}_{kd}=0\end{aligned} \tag{30}$$

This set has explicit solutions given by

$$\begin{aligned}\grave{\sigma}_{kd}&=\left(\frac{\overleftarrow{C}_{kd}}{\tilde{C}_k}\left(1+\left(\frac{\overrightarrow{C}_{kd}}{\overleftarrow{C}_{kd}}\right)^{1/3}\right)\right)^{1/2}\\ \acute{\sigma}_{kd}&=\left(\frac{\overrightarrow{C}_{kd}}{\tilde{C}_k}\left(1+\left(\frac{\overleftarrow{C}_{kd}}{\overrightarrow{C}_{kd}}\right)^{1/3}\right)\right)^{1/2}\end{aligned} \tag{31}$$

The emerging algorithm consists of repeated iterations, where at each iteration, the known 'old' parameters Λ^o are updated into 'new' parameters Λ^v.

E-step

(I) Given the training data $\mathcal{X}=\{\boldsymbol{x}_t, t=1, \ldots, T\}$ and the model $\Lambda^o=\left\{w_k^o, \boldsymbol{\xi}_k^o, \grave{\boldsymbol{\sigma}}_k^o, \acute{\boldsymbol{\sigma}}_k^o, k=1, \ldots, K\right\}$, calculate $P(k/\boldsymbol{x}_t, \Lambda^o)$ using (21) for $k=1, \ldots, K$ and $t=1, \ldots, T$.

M-step

(II) For $k=1, \ldots, K$, calculate new weights w_k^v using (18).
(III) For $d=1, \ldots, D$ and $k=1, \ldots, K$, calculate new ξ_{kd}^v using (23) and assuming $\Lambda^o=\left\{w_k^v, \xi_k^o, \grave{\boldsymbol{\sigma}}_k^o, \acute{\boldsymbol{\sigma}}_k^o, k=1, \ldots, K\right\}$.
(IV) For $d=1, \ldots, D$ and $k=1, \ldots, K$, calculate the three auxiliary coefficients $\tilde{C}_k$, $\overleftarrow{C}_{kd}$ and $\overrightarrow{C}_{kd}$ in (27)–(29) assuming $\Lambda^o=\left\{w_k^v, \xi_k^v, \grave{\boldsymbol{\sigma}}_k^o, \acute{\boldsymbol{\sigma}}_k^o, k=1, \ldots, K\right\}$. Then solve (30) to obtain $\grave{\sigma}_{kd}^v$ and $\acute{\sigma}_{kd}^v$.
(V) Rename the completed new set from 'new' to 'old', $\left\{w_k^v, \xi_k^v, \grave{\boldsymbol{\sigma}}_k^v, \acute{\boldsymbol{\sigma}}_k^v\right\}\rightarrow\left\{w_k^o, \xi_k^o, \grave{\boldsymbol{\sigma}}_k^o, \acute{\boldsymbol{\sigma}}_k^o\right\}$, for all k, and go to step I.

Steps I–V are repeated till convergence.

4 Speaker recognition experiments

We performed some experiments in order to examine the above EM training algorithm and compare the performance of Gaussian and skew GMMs in some basic speaker recognition tasks. Models were trained using MFCC and LSF as feature vectors. The preprocessing involved only pre-emphasis and removal of silent frames. For training, the EM and the skew EM algorithms were used. Initial values of the Gaussians means, and the location parameters ξ of the skew Gaussians, were set as the centres of the clusters calculated by the $\boldsymbol{K}$-means algorithm (starting with $\boldsymbol{K}$ random vectors from the data set $\mathcal{X}$ and improving them iteratively using Euclidean distance). The initial values of σ were obtained using the standard deviation in each cluster. For the skew Gaussian, the initial values were assigned as $\grave{\sigma}=\sigma$ and $\acute{\sigma}=10*\sigma$. The same preprocessing and feature extraction were used both in training and testing. The probability of each speaker to utter the test file was calculated, normalised and will be reported by detection error trade-off (DET) curves for the clean speech and for the noisy speech only by equal error rate (EER) tables to save space.

4.1 Setting

The first set of experiments was conducted using the TIMIT speech database [18, 19]. This database provides speech of 8 kHz bandwidth, no intersession variability, no acoustic noise, no microphone variability and channel distortion. Consequently, the relative recognition performance can be fully attributed to the models and the features that were used. Afterward all the experiments were repeated with the NTIMIT database. The NTIMIT database presents a replica of the TIMIT database with speech degraded by carbon-button transduction and its transmission of the speech through telephone lines with varying conditions [20]. Performance comparison between identical experiments with TIMIT and NTIMIT examines the impact of telephone transmission degradations on the modelling method.

Our testing approach followed the 'long training/short testing' protocol, suggested by Bimbot *et al.* [21] for speaker recognition using TIMIT database and adapted to NTIMIT. The 'long training' comprised five sentences with an average total duration of a 14.4 s and ‘short training’ was composed of two sentences with an average total duration of 5.7 s. The ‘long testing’, included five sentences lasting about 15.9 s. The ‘short testing’ included two sentences lasting 3.2 s.

The skew GMM requires about 50% more parameters than the symmetric GMM (because a skew Gaussian involves ξ, $\grave{\sigma}$ and $\acute{\sigma}$ as opposed to just mean and standard deviation for the symmetric Gaussian). In order to have approximately an equal overall number of parameters for a fair comparison of performance, we compared skew GMMs with $K_{\text{skew}} = 12$ components to symmetric GMMs with $K_{\text{sym}} = 18$ components in one enrolment and $K_{\text{skew}} = 24$ to $K_{\text{sym}} = 36$ in a second enrolment. In all the experiments the feature vectors (LSF or MFCC) were of size $D = 13$.

Closed set speaker recognition experiments were held using 100 speakers selected randomly from the database. Per each test scenario 10 000 combinations were measured, where in 100 of them, true speaker was evaluated by his own model and in the rest by models of other speakers.

4.2 Results

4.2.1 TIMIT results: Results of the experiments with the TIMIT database are presented by DET graphs of the percent of approved incorrect speakers (miss probability) against the number of rejected correct speakers (false alarm) created by varying the threshold for acceptance/rejection.

Fig. 4 presents the performance of GMMs and skew GMMs with LSF and MFCC as feature vectors, using short training with short testing (in the upper graph) and long testing (in the lower graph). The number of mixture components were 12 for the skew GMM and 18 for the symmetric GMM. In the ‘short-short’ scenario, that uses very little amounts of data for training and testing, the performance is poor for all models. However, the skew GMMs with LSF are seen to be better by more than 10% than all the other models. Examining next the ‘short-long’ scenario (the lower graph), the skew GMMs with LSF and MFCC are superior to all other models by more than 5%. The overall performance when using small amount of data for training is still unsatisfactory, but again the performances of the skew GMMs are better. Note also how the increase in testing data, significantly improves the scores of the ‘short–long’ results in comparison with the ‘short–short’ results.

Fig. 5 brings corresponding results with long training experiments with short testing in the upper figure and long testing in the lower figure with again 12 skew Gaussians and 18 symmetric Gaussians. As seen, the increased amount of training data improves the performance of all the results in the ‘long–short’ test compared to the ‘short–short’ case. However, the skew GMMs with LSF are still somewhat better then the others. The lower part of Fig. 5 brings the results of ‘long–long’ experiments. Here, both skew and symmetric speaker models perform well and the edge of skew Gaussians over the symmetric Gaussians becomes insignificant (1%).

In a second round of experiments, we repeated the above experiments using skew GMMs with 24 Gaussians and symmetric GMMs with 36 Gaussians. Fig. 6 is the ‘24–36’ correspondent of the ‘12–18’ results in Fig. 4. Similarly, Fig. 7 corresponds to Fig. 5. Generally speaking, one can

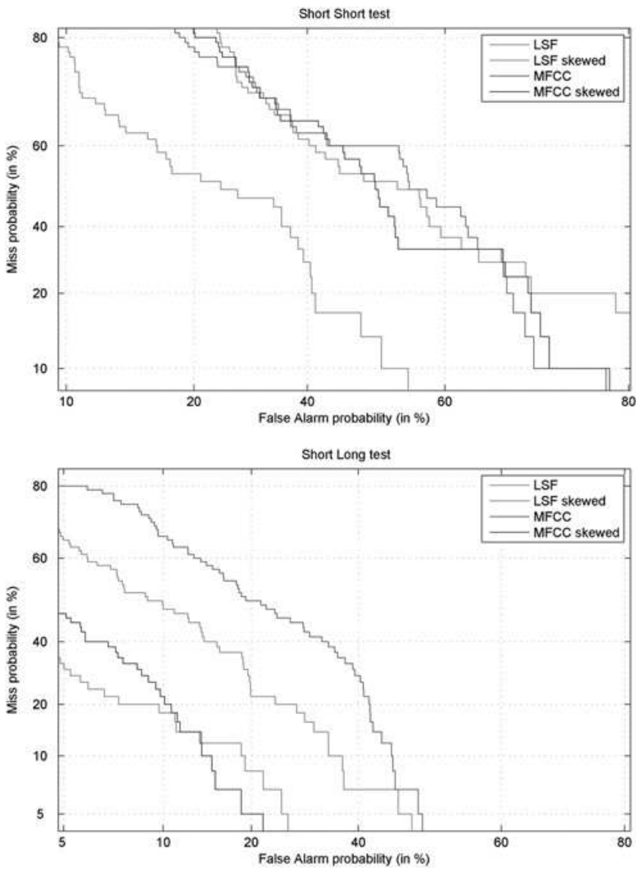

Fig. 4 *DET curves for short training with short testing (above) and long testing (below); $K_{skew} = 12$, $K_{sym} = 18$*

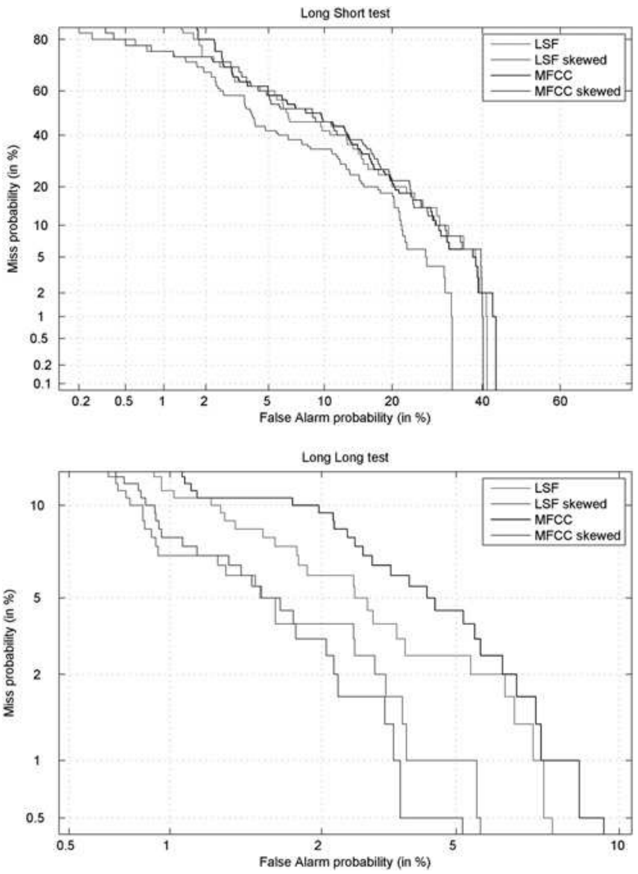

Fig. 5 *DET curves for long training with short testing (above) and long testing (below); $K_{skew} = 12$, $K_{sym} = 18$*

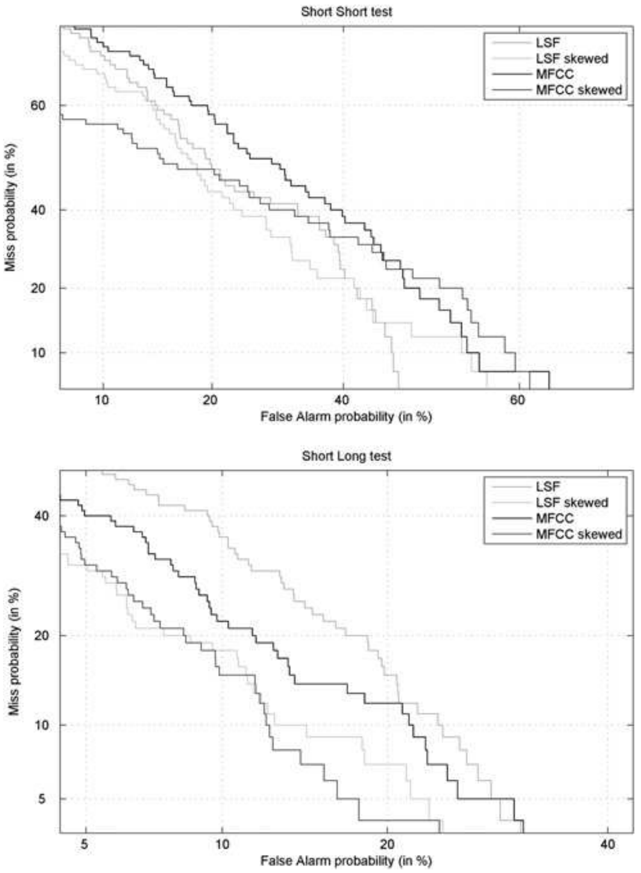

Fig. 6 *DET curves for short training with short testing (above) and long testing (below); K_{skew} = 24, K_{sym} = 36*

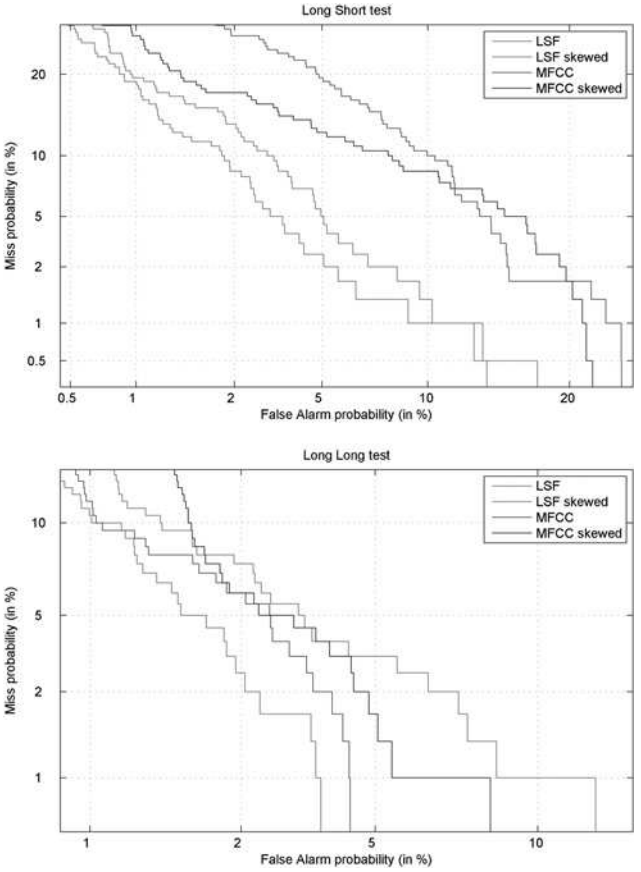

Fig. 7 *DET curves for short training with short testing (above) and long testing (below); K_{skew} = 24, K_{sym} = 36*

Table 1 Noisy speech EER scores using $K_{\text{skew}} = 12$ and $K_{\text{sym}} = 18$

EER	Short–short	Short–long	Long–short	Long–long
LSF	50	48	32	26
LSF skewed	48	46	25	16
MFCC	55	44	26	23
MFCC skewed	51	47	31	24

Table 2 Noisy speech EER scores using $K_{\text{skew}} = 24$ and $K_{\text{sym}} = 36$

EER	Short–short	Short–long	Long–short	Long–long
LSF	51	48	31	24
LSF skewed	47	44	25	18
MFCC	55	42	30	22
MFCC skewed	49	48	30	24

see that scores improved as the size of the models doubled. Otherwise, the trends remain essentially similar in comparing performance of skew GMMs and symmetric GMMs and comparing LSF to MFCC. In the short-long scenario (lower graph in Fig. 6) the skew GMMs with either LSF or MFCC are more than 5% better than the symmetric GMMs. In the 'long-long' scenario (lower graph in Fig. 7) both skew and symmetric speaker models perform equally well (EER under 5%) but still the skew models (with both LSF and MFCC) are somewhat better than the symmetric models.

4.2.2 NTIMIT results: We repeated all the above reported experiments with the NTIMIT data base [20]. For brevity, results in this section are brought only by their EER values. Table 1 compares the performance of LSF and MFCC with short and long training using the smaller order GMMs ($K_{\text{skew}} = 12$ and $K_{\text{sym}} = 18$). The performance of all models deteriorates for the current noised speech. Note the impact of the amount of training and testing data. The 'long' training and testing experiments perform better than the 'short' training experiments for both feature vectors and model types. The lowest EER (of 16%) was attained with skew GMM using LSF.

Table 2 brings results of corresponding experiments using the higher order GMMs ($K_{\text{skew}} = 24$ and $K_{\text{sym}} = 36$). The performance remains poor. Again the 'long–long' experiments are better than the rest again the best performer is the skew GMM with LSF (attaining an EER of 18%).

4.3 Discussion

Our primary goal in the above experiments was to explore the relative modelling capabilities of GMM and skew GMM through their scores in speaker identification experiments. When analysing the performance of skew GMM and GMM, one can see that skew GMM outperforms the standard GMM by a few percents in most of the test scenarios. This observation becomes more prominent when looking at the modelling capabilities with small amounts of clean data (TIMIT), as in the 'short–short' test scenario presented in Fig. 4. In this case skew Gaussian models with LSF outperform all the other models by over 10%. This is also true for the 'short–long' part of Fig. 4 where skew GMMs with both LSF and MFCC outperform the other models.

doi: 10.1049/iet-spr.2013.0270

Similar results can be seen when looking at the performance of the larger models in Fig. 6. This indicates that the skew GMM can capture better the skewness of speech feature vectors, thus providing better discrimination than the symmetric GMM. As additional training data becomes available, the difference between the skew and the symmetric models diminishes.

It is also interesting to compare the performance of identical sets of models and systems, with the same amount of data, when the only difference is the quality of the database (TIMIT against NTIMIT). The reduced quality of speech caused a significant performance degradation throughout. In the 'long' training scenarios, a performance degradation of about 20% can be seen in the 'long–long' tests and 20–30% in the 'long–short' tests in transition from TIMIT to NTIMIT. However, the degradation of the skew GMM is lower than the degradation of GMM in all test scenarios. This observation suggests that the skew GMM is more resilient to adverse conditions than the symmetric GMM.

5 Concluding remarks

The paper presented an EM algorithm for training skew GMMs and examined it by performing some basic speaker recognition experiments. The skew and the symmetric GMMs were trained with two types of feature vectors, the widely used (in speech and speaker recognition) MFCC, and with LSF that (together with the similarly distributed ISF) are used in most speech coding standards and exhibit significantly more skewed distribution than MFCC.

These experiments were designed mostly to validate the training algorithm and examine the capacity of a mixture of skew Gaussians to fit skew distributed data. They used a relatively low-order models and short feature vectors that were trained and tested with a limited amount of data using a standard setting of the TIMIT and NTIMIT databases. The LSF achieved better performance with skew GMMs than with symmetric GMMs. Actually, the skew GMMs with LSF outperformed both skew and symmetric GMMs with MFCC. When repeating the experiments with noisy speech, the performance degraded significantly but the skew GMM demonstrated better resilience to the deterioration of the speech.

More research is required in order to determine to what extent higher scale speaker and speech recognition applications would benefit from using skew GMMs. It is possible that the relative advantage of skew Gaussians over symmetric Gaussians decreases as the overall number of model parameters is increased, leaving its use attractive mostly for tasks that inherently have access to only a limited amount of data (as in e.g. forensic scenarios). We also observed (in some not yet conclusive experiments held toward wrapping up this paper for publication) that we can improve recognition scores by choosing initial model parameters slightly differently. Consequently, further study of new ways to initiate the suggested EM algorithm is also required. The improved performance of LSF with skew GMMs makes them of interest in the context of speech transmission standards. The LSF or ISF that are used there to compress speech might be used to recognise speakers directly from the raw transmitted parameters inside the coders. In considering the replacement of GMMs with skew GMMs in speech applications it is worth noting that the training and testing of skew GMMs is not truly more complicated and that skew Gaussians admit the symmetric Gaussians as a special case.

6 References

1 Mclachlan, G., Peel, D.: 'Finite mixture models' (John Wiley and Sons Inc., 2000)

2 Reynolds, D.A.: 'Speaker identification and verification using Gaussian mixture speaker models', *Speech Commun.*, 1995, **17**, pp. 91–108

3 Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: 'Speaker verification using adapted Gaussian mixture models', *Digit. Signal Process.*, 2000, **10**, pp. 19–41

4 Kleijn, W.B., Paliwal, K.K. (Ed): 'Speech coding and synthesis' (Elsevier Science, Amsterdam, The Netherlands, 1995)

5 ITU-T Recommendation G.718: 'Frame error robust narrow-band and wideband embedded variable bit-rate coding of speech and audio from 8–32 kbit/s', 06/2008

6 Jelinek, M., Salami, R.: 'Wideband speech coding advances in VMR-WB standard', *IEEE Trans. Audio Speech Language Process.*, 2007, **15**, pp. 1167–1179

7 Bistritz, Y., Peller, S.: 'Immittance Spectral Pairs (ISP) for speech encoding'. Proc of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Minneapolice, Minnesota, April 1993, vol. 2, pp. 9–12

8 Zilca, R., Bistritz, Y.: 'Distance based Gaussian Mixture Model for speaker recognition over the telephone'. Proc. of the 6th Int. Conf. on Spoken Language Processing, Beijing, China, October 2000, pp. 16–20

9 Lee, B.J., Kim, S., Kang, H.G.: 'Speaker recognition based on transformed line spectral frequencies'. Proc. of IEEE Intelligent Signal Processing and Communication Systems, November 2004, ISPACS'04, pp. 177–180

10 Cordeiro, H., Ribeiro, C.M.: 'Speaker characterization with MLSFs'. Speaker and Language Recognition Workshop, IEEE Odyssey 2006, San Juan, Puerto Rico, June 2006, pp. 1–4

11 Matza, A., Bistritz, Y.: 'Skew Gaussian mixture models for speaker recognition'. Proc of 12th Annual Conf. of Int. Speech Communication Association, Florence, Italy, August 2011, pp. 5–8

12 Gibbons, J.F., Mylroie, S.: 'Estimation of impurity profiles in ion-implanted amorphous targets using joined half-Gaussian distributions', *Appl. Phys. Lett.*, 1973, **22**, pp. 568–569

13 John, S.: 'The three-parameter two-piece normal family of distributions and its fitting', *Commun. Stat. – Theory Methods*, 1982, **11**, (8), pp. 879–885

14 Azzalini, A.: 'A class of distributions which includes the normal ones', *Scand. J. Stat.*, 1985, **12**, pp. 171–178

15 Azzalini, A.: 'Further results on a class of distributions which includes the normal ones', *Statistica*, 1986, **XLVI**, pp. 199–208

16 Arellano-Valle, R.B., Gomez, H.W., Quintana, F.A.: 'Statistical inference for a general class of asymmetric distributions', *J. Statistical Planning Inference*, 2005, **128**, (2), pp. 427–443

17 Bilmes, J.A.: 'A gentle tutorial of the EM algorithm and its appliation to parameter estimation for Gausian mixtures and hidden Markov models', Technical Report TR-97–021 EECS U.C. Berkley, 1998

18 Campbell Jr., J.P., Reynolds, D.A.: 'Corpora for the evaluation of speaker recognition systems'. Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Phoenix, Arizona, May 1999, pp. 2247–2250

19 Zue, V., Seneff, S., Glass, J.: 'Speech database development: TIMIT and beyond'. ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases, Noordwijkerhout, the Netherlands, September 1989, pp. 20–23

20 Fisher, W.M., Doddington, G.R., Goudie-Marshall, K.M., *et al.*: 'NTIMIT – LDC Catalog No. LDC93S2', http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC93S2

21 Bimbot, F., Magrin-Chagnolleau, I., Mathan, L.: 'Second-order statistical measures for text-independent speaker identification', *Speech Commun.*, 1995, **17**, (1), pp. 177–192