Skew Gaussian mixture models for speaker recognition

Avi Matza and *Yuval Bistritz*

School of Electrical Engineering
Tel-Aviv University, Tel-Aviv 69978, Israel
avimatza@eng.tau.ac.il, bistritz@eng.tau.ac.il

Abstract

The current paper proposes skew Gaussian mixture models for speaker recognition and an associated algorithm for its training from experimental data. Speaker identification experiments were conducted, in which speakers were modeled using the familiar Gaussian mixture models (GMM), and the new skew-GMM. Each model type was evaluated using two sets of feature vectors, the mel-frequency cepstral coefficients (MFCC), that are widely used in speaker recognition applications, and line spectra frequencies (LSF), that are used in many low bit rate speech coders but were not that successful in speech and speaker recognition. Results showed that the skew-GMM, with LSF, compares favorably with the GMM-MFCC pair (under fair comparison conditions). They indicate that skew-Gaussians are better suited for capturing the relatively highly non-symmetrical shapes of the LSF distribution. Thus the skew-GMM with LSF offers a worthy alternative to the GMM-MFCC pair for speaker recognition.

Index Terms: Speaker recognition, Gaussian mixture models, skew-Gaussians, line spectral frequencies.

1. Introduction

Most speaker identification and verification systems today use Gaussian mixture models (GMM) with mel frequency cepstral coefficients (MFCC) as feature vectors [1, 2]. There are several good reasons for using GMM. For one is the fact that a linear combination of Gaussians can effectively model complex distributions. Another reason is the central limit theorem, which states that the average of a sufficiently large number of independent random variables, each with finite mean and variance, will be approximately normally distributed. Naturally, a GMM model matches best empirical distributions whose shape is a linear combination of several symmetric bell shaped curves. The symmetric empirical distribution of the mel cepstra parameters supports this requirement quite well, as can be realized from Figure. 1. However, other feature vectors, with less symmetric distribution characteristics, are less likely to perform well with standard GMM, even if otherwise they could be argued to carry good discrimination power. Interesting examples for such features are line spectral frequencies (LSF) (also known as line spectral pairs) [3], used in CELP based coder standards e.g. [4], or immittance spectral pairs (ISP) [5] used in more recent standards e.g. [6]. Since these features demonstrate good correlation with the location of the formants in voiced speech, they could be argued to offer good discrimination capabilities between phonemes and speakers. Yet, attempts to use LSF in speaker recognition applications, well justified by their role in speech coders, were not too rewarding, [7, 8, 9]. When exploring the histogram structure of LSF, see Figure 2, it is apparent that the LSF are much more skewed than the MFCC.

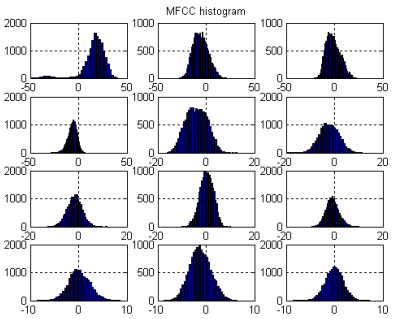

Figure 1: *Histograms of 12 MFCC coefficients*

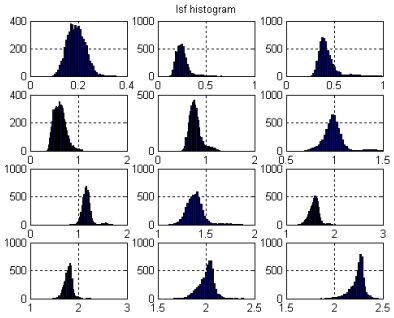

Figure 2: *Histograms of 12 LSF coefficients*

It is possible to use higher order GMMs to model empirical distributions that have asymmetrical properties. However, additional Gaussians inherently increase model variance, requiring more data for proper training and will not necessarily increase performance. An alterative approach might be replacing the symmetrical Gaussian distribution with a more skewed distribution. It is reasonable to expect that skewed distributions will capture asymmetrical distributions better and will require less training data when modeling features like LSF. In this paper we explore this approach and replace GMMs with a skewed-Gaussian mixture models. There are, of course, other known skewed distributions that could be explored to this end, but we chose skew Gaussians for two reasons. One is that skew-GMM offers a soft deviation from GMM that admits the symmetric bell shape when it best suits the training data. The second is that the training algorithm complexity of the skew-GMM model, is comparable to the traditional GMM training.

The paper is constructed as follows. The skew Gaussian distribution and its use in a mixture for speaker modeling are

presented in the next section. Section 3 describes the experimental setting followed by a presentation of recognition results followed by a short discussion and conclusions.

2. Skew Gaussian mixture models

The skew-Gaussian distribution probability density functions, that was proposed by Azzalini in [10, 11], has various representations. Let $\phi(x) \sim \mathcal{N}(0,1)$ denote the standard Gaussian probability density function (pdf) for a scalar random variable x, the skew-Gaussian pdf can be defined as

$$a(x;\xi,\grave{\sigma},\acute{\sigma}) = \frac{2}{(\acute{\sigma}+\grave{\sigma})}\left(\phi\left[\frac{x-\xi}{\acute{\sigma}}\right]u(x-\xi)+\phi\left[\frac{x-\xi}{\grave{\sigma}}\right]u(\xi-x)\right) \quad (1)$$

where $u(x)$ denotes the unit step function ($u(x)=1$ for $x \geq 0$ and 0 for $x < 0$), ξ is the location parameter, $\grave{\sigma}$ is the left shape parameter and $\acute{\sigma}$ is the right shape parameter. We shall use $x \sim \mathcal{SN}(\xi,\grave{\sigma},\acute{\sigma})$ to say that the pdf of the random variable x is given by (1).

The generalization of the skew Gaussian pdf to a multivariate vector $\boldsymbol{x} = [x_1,\ldots,x_D]^t$ will be denoted by $S(\boldsymbol{x})$. For independent $x_d \sim \mathcal{SN}(\xi_d,\grave{\sigma}_d,\acute{\sigma}_d)$, $d=1,...,D$ it becomes

$$S(\boldsymbol{x}) = S(\boldsymbol{x};\boldsymbol{\xi},\grave{\boldsymbol{\sigma}},\acute{\boldsymbol{\sigma}}) = \prod_{d=1}^{D} a(x_d;\xi_d,\grave{\sigma}_d,\acute{\sigma}_d), \quad (2)$$

and we shall indicate this case by $\boldsymbol{x} \sim \mathcal{SN}(\boldsymbol{\xi},\grave{\boldsymbol{\sigma}},\acute{\boldsymbol{\sigma}})$. For better acquaintance with skew Gaussian distribution and its properties consider [12].

A skew Gaussian mixture model of order K, associate a vector $\boldsymbol{x}$ of independent random variables, with a weighted sum of multivariate skew-Gaussians

$$P(\boldsymbol{x}|\lambda) = \sum_{k=1}^{K} w_k S_k(\boldsymbol{x}) \quad (3)$$

where $S_k(\boldsymbol{x}) = S(\boldsymbol{x};\boldsymbol{\xi}_k,\grave{\boldsymbol{\sigma}}_k,\acute{\boldsymbol{\sigma}}_k)$ and $w_k > 0$, $k=1,...,K$ are the weights, such that $\sum_{k=1}^{K} w_k = 1$. λ is used to denote the collection of parameters that define the Skew-GMM, $\lambda := \{w_k,\boldsymbol{\xi}_k,\grave{\boldsymbol{\sigma}}_k,\acute{\boldsymbol{\sigma}}_k, k=1,...K\}$. The parameter vectors $\boldsymbol{\xi}_k,\grave{\boldsymbol{\sigma}}_k,\acute{\boldsymbol{\sigma}}_k$ are vectors of length D with entries denoted by $\xi_{kd},\grave{\sigma}_{kd},\acute{\sigma}_{kd}, k=1,...D$

Next we need a procedure to infer the parameters of a skew-GMM from empirical data, in order to train speaker models. Since the training of mixture model optimization involves maximization of log of sums of pdfs, that is not tractable to to analytical optimization, numerous algorithms were suggested for this end. The most popular method for training GMM is the Expectation-Maximization (EM) algorithm that was proved to converge to the maximum log-likelihood estimator [13]. The likelihood that a set of feature vectors $X = \{\boldsymbol{x}_1,\boldsymbol{x}_2\ldots,\boldsymbol{x}_T\}$ stems from a model with parameters λ, is defined by the probability $Pr(X|\lambda)$. Assuming that the vectors are statistically independent $Pr(X|\lambda) = \prod_{t=1}^{T} P(\boldsymbol{x}_t|\lambda)$. The validity of the assumption that the collection of vectors X belongs to the skew-GMM model is measured by its log likelihood,

$$L = \log\left[Pr(X|\lambda)\right] = \sum_{t=1}^{T} \log\left[\sum_{k=1}^{K} w_k S_k(\boldsymbol{x}_t)\right] \quad (4)$$

The training of the Skew-GMM toward maximizing this log likelihood can be done by the next presented EM iterations (whose derivation will be shown elsewhere). At each iteration there is an available Skew-GMM model denoted λ^o (with superscript ^{o} for all its parameters) that is updated to a next λ^ν (with superscript $^\nu$ for all its parameters) using the training data $X = \{\boldsymbol{x}_1,\boldsymbol{x}_2\ldots,\boldsymbol{x}_T\}$. In the following we also use $x_{td}, d=1,...D$ to denote the entries of the observed vector $\boldsymbol{x}_t$.

E-step

For $k=1,...,K$ and for $t=1,...,T$ calculate:

$$P_{kt}(k|\boldsymbol{x}_t,\lambda^o) = \frac{w_k^o S_k(\boldsymbol{x}_t)}{\sum_{k=1}^{K} w_k^o S_k(\boldsymbol{x}_t)} = \frac{w_k^o S_k(\boldsymbol{x}_t)}{P(\boldsymbol{x}_t|\lambda^o)}\,, \quad (5)$$

M-step

I) For $k=1,...,K$ calculate the next weights w_k^ν by

$$w_k^\nu = \frac{1}{T}\sum_{t=1}^{T} P_{kt}(k \mid \boldsymbol{x}_t,\lambda^o) \quad (6)$$

II) Set $\lambda^\nu = \{w_k^\nu,\boldsymbol{\xi}_k^o,\grave{\boldsymbol{\sigma}}_k^o,\acute{\boldsymbol{\sigma}}_k^o, k=1,...K\}$. For $d=1,...,D$ and for $k=1,...,K$ calculate:

$$\xi_{kd}^\nu = \frac{\sum_{t=1}^{T} x_{td} P_{kt}(k|x_{td},\lambda^\nu)u(x_{td}-\xi_{kd}^o)}{\sum_{t=1}^{T} P_{kt}(k|x_{td},\lambda^\nu)u(x_{td}-\xi_{kd}^o)} + \frac{\sum_{t=1}^{T} x_{td} P_{kt}(k|x_{td},\lambda^\nu)u(\xi_{kd}^o-x_{td})}{\sum_{t=1}^{T} P_{kt}(k|x_{td},\lambda^\nu)u(\xi_{kd}^o-x_{td})} \quad (7)$$

III) Set $\lambda^\nu = \{w_k^\nu,\boldsymbol{\xi}_k^\nu,\grave{\boldsymbol{\sigma}}_k^o,\acute{\boldsymbol{\sigma}}_k^o, k=1,...K\}$. For $k=1,...,K$ and for $d=1,...,D$ calculate the following three auxiliary terms:

$$\widetilde{C}_{kd} = \sum_{t=1}^{T} P_{kd}(k|x_{td},\lambda^\nu)$$

$$\overleftarrow{C}_{kd} = \sum_{t=1}^{T} (x_{td}-\xi_{kd}^\nu)^2 P_{kd}(k|x_{td},\lambda^\nu)u(\xi_{kd}^\nu-x_{td})$$

$$\overrightarrow{C}_{kd} = \sum_{t=1}^{T} (x_{td}-\xi_{kd}^\nu)^2 P_{kd}(k|x_{td},\lambda^\nu)u(x_{td}-\xi_{kd}^\nu)$$

Then, obtain $\grave{\sigma}_{kd}^\nu$ and $\acute{\sigma}_{kd}^\nu$ by solving the following pair of equations.

$$\widetilde{C}_{kd}(\grave{\sigma}_{kd}^\nu)^3 - \overleftarrow{C}_{kd}\grave{\sigma}_{kd}^\nu - \overleftarrow{C}_{kd}\acute{\sigma}_{kd}^\nu = 0 \quad (8)$$

$$\widetilde{C}_{kd}(\acute{\sigma}_{kd}^\nu)^3 - \overrightarrow{C}_{kd}\acute{\sigma}_{kd}^\nu - \overrightarrow{C}_{kd}\grave{\sigma}_{kd}^\nu = 0 \quad (9)$$

IV) Complete $\lambda^\nu = \{w_k^\nu,\boldsymbol{\xi}_k^\nu,\grave{\boldsymbol{\sigma}}_k^\nu,\acute{\boldsymbol{\sigma}}_k^\nu, k=1,...,K\}$.

V) Replace $\lambda^o \leftarrow \lambda^\nu$ and repeat the E-M steps until convergence.

3. Speaker recognition experiments

We ran speaker recognition experiments in order to compare the performance of Gaussian and skew-Gaussian mixture models. Since the purpose of the comparison was to evaluate the relative performance of the two models and in order to minimize the impact of non relevant side effects, the system we used was kept as simple as possible. The models were trained using raw MFCC and LSF without further pre processing. The initial estimation of mixture parameters was generated through the application of the K-means algorithm. The same processing and feature extraction was used both in training and testing. The probability of each speaker to utter the test file was calculated, and the file was attributed to the speaker with the highest scoring model.

3.1. The setting

The experiments were conducted using the TIMIT speech database [14, 15]. This database is most appropriate for our purpose since it enable us to examine speaker recognition system performance under almost ideal conditions. It has 8 kHz bandwidth without intersession variability, no acoustic noise, no microphone variability and distortions etc., thus recognition errors can be attributed to the used model. Our testing approach followed the "long training / short test" protocol, suggested by Bimbot [16] for speaker recognition using TIMIT database. This testing protocol was designed for speaker recognition systems with closed set testing, and its purpose was to allow a methodical comparison between various speaker recognition systems. The idea was to have two types of training sessions and two types of testing sessions and inspect system performance with various combinations. The "long training" composed of 5 sentences with an average total duration of a 14.4 seconds and "short training" was composed of 2 sentences having an average total duration of 5.7 seconds. The "long test", was also 5 sentences long and lasts about 15.9 seconds where the "short test" was 2 sentences long and 3.2 seconds in average. For a short review of various speaker recognition systems performance using TIMIT database one could look at Bimbot [17]. Experiments were held with 10,25,50 and 100 speakers selected randomly from the database.

The skew-Gaussian pdf involves more parameters than the standard Gaussian (as it replaces the σ with the pair $\grave{\sigma}$ and $\acute{\sigma}$). Therefore, for a fair evaluation (with approximately the same overall number of parameters) GMM and Skew-GMM should be compared with different number of mixture components (K). To meet this requirement, we used Skew-GMM systems with $K = 12$ components and GMM systems with $K = 18$. All the experiments used LSF and MFCC feature vectors of size $D = 22$.

3.2. The results

Following are recognition results from our speaker identification experiments using GMM and Skew-GMM. The results were obtained from the two models with both MFCC and LSF as feature vectors, within the settings described above.

Figure 3 demonstrates the performance of GMM and Skew-GMM models, using LSF and MFCC feature vectors with short training and short testing. It can be seen that MFCC performance are rapidly decreasing, while skewed LSF has almost constant performance. As a result, Skew-GMM with LSF is significantly better than all the other models and for 100 speakers it is better by 15% than the others. One can also observe that the Skew-GMM with MFCC does not perform better than

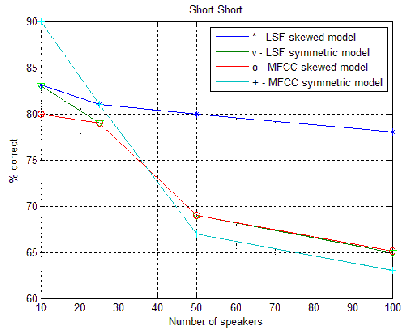

Figure 3: *Recognition in short-short tests*

GMM with MFCC. Apparently in this case, the Skew-GMM system is unable to gain the additional performance it attains for the skewed LSF, due to the fact that MFCC is considerably less skewed than LSF. We can conclude that when only small amounts of reliable data exists, Skew-GMM with LSF performs much better than all other systems.

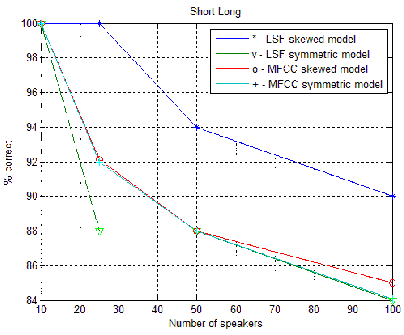

Figure 4: *Recognition in short-long tests*

Figure 4 considers the "short-long" test scenario. It shows that Skew-GMM with LSF is superior to all others by approximately 5%. Again, recognition results of the skewed model with MFCC achieve results comparable to the non skewed model with MFCC. Notice also how the additional testing information, significantly increases the scores of "short-long" test (Figure 4) in comparison with the "short-short" results (Figure 3).

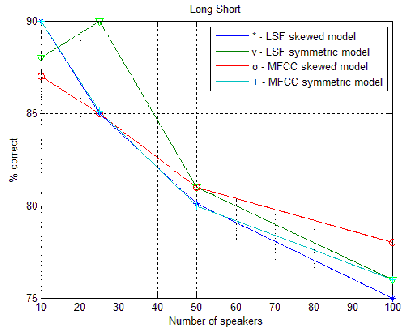

Figure 5: *Recognition in long-short tests.*

Figure 5 brings the results of the "long-short" experiments. It can be seen that as a result of the increased amount of training data, all systems demonstrate similar performance.

Finally, Figure 6 presents the results obtained in "long-long" experiments. We can see that both Skew-GMM and GMM speaker models perform very well. Again, the edge of skewed modeling over the symmetric Gaussians becomes insignificant (just 1% at 100 speakers). In summary, Skew-GMM with LSF maintains a "better or equal" performance when compared to GMM with MFCC in all our testing scenarios, but the differences tend to decrease as the available training data increases.

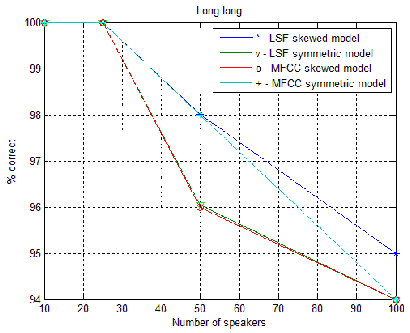

Figure 6: *Recognition in long-long tests*

3.3. Discussion

When comparing system performance using Skew-GMM and GMM for speaker modeling, 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 using small amounts of data for training, like in the "short-short" test scenario of Figure 3, were skew Gaussian model outperforms all other models by as much as 15%. While skew-GMM improves LSF performance significantly compared to GMM, it does not have too much to offer to the relatively symmetrically distributed MFCC. The ability of the skew-GMM to capture better the asymmetry of the LSF distribution, boosts the discrimination capability of the skew-GMM with LSF to a level that is competitive with the GMM-MFCC pair.

4. Conclusions

The paper proposed skew Gaussian mixture models for speaker recognition. The Skew-GMM model includes the GMM as a special case, and it can be trained by a reasonably simple extension of its EM training. In experiments held to evaluate the new approach, Skew-GMM with the asymmetrically distributed LSF outperformed comparable systems based on GMM with MFCC.

5. References

[1] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", *Speech Communication*, vol. 17, pp. 91-108, 1995.

[2] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification", *IEEE signal processing letters*, vol. 13 (5), 2006.

[3] W. B. Kleijn and K. K. Paliwal, Editors, "Speech Coding And Synthesis", Elsevier Science, Amsterdam, The Netherlands (1995).

[4] 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.

[5] Y. Bistritz and S. Peller, "Immittance Spectral Pairs (ISP) for speech encoding", *Proc of the 1993 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, ICASSP '93 vol. 2 , pp. 9-12, April 1993, Minneapolis, Minnesota.

[6] M. Jelinek and R. Salami, "Wideband speech coding advances in VMR-WB standard", *IEEE Transactions on Audio, Speech and Language Processing*, vol 15 pp. 11671179, May 2007.

[7] R. Zilca and Y. Bistritz, "Distance based Gaussian Mixture Model for speaker recognition over the telephone" *Proc. of the 6th International Conference on Spoken Language Processing*, ICSLP 2000, pp. 16-20, October 2000, Beijing, China.

[8] B. J. Lee, S. Kim, H-G. Kang, "Speaker recognition based on transformed line spectral frequencies", *Proc. of IEEE Intelligent Signal Processing and Communication Systems* ISPACS'04, November 2004, pp. 177-180

[9] H. Cordeiro, C. M. Ribeiro, "Speaker characterization with MLSFs", *Speaker and Language Recognition Workshop, IEEE Odyssey 2006*, pp. 1-4, June 2006, San Juan, Puerto Rico.

[10] A. Azzalini, "A class of distributions which includes the normal ones", *Scand. J. Statist.* vol. 12, pp. 171-178, 1985.

[11] A. Azzalini, "Further results on a class of distributions which includes the normal ones", *Statistica*, vol. XLVI, pp. 199-208, 1986.

[12] T.I. Lin, J. C. Lee and S. Y. Yen, "Finite mixture modelling using the skew normal distribution", *Statistica Sinica* vol. 17, pp. 909-927, 2007.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm", *J. Royal Statist. Soc Ser. B.*, vol. 39, 1977.

[14] J. P. Campbell, Jr. and D. A. Reynolds, "Corpora for the evaluation of speaker recognition systems", *Proc. of the 1999 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* , ICASSP '99, pp. 2247-2250, May 1999, Phoenix, Arizona.

[15] V. Zue, S. Seneff and J. Glass, "Speech database development: TIMIT and beyond", *ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases*, pp. 20-23, September 1989, Noordwijkerhout, The Netherlands.

[16] F. Bimbot, I. Magrin-Chagnolleau and L. Mathan, "Second-order statistical measures for text-independent speaker identification", *Speech Communication*, vol. 17 (1) pp. 177-192, August 1995.

[17] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification", *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430-451, 2004.