Pitch and MFCC dependent GMM models for speaker identification systems

*Hassan Ezzaidi and Jean Rouat**

Ermetis, Université du Québec à Chicoutimi, Chicoutimi, Québec, Canada, G7H 2B1
* IMSI, Université de Sherbrooke, Sherbrooke, Québec, Canada, J1K 2R1
hezzaidi@uqac.uquebec.ca, Jean.Rouat@ieee.org

Abstract

Raising the performance of the systems identification speaker still constitutes the object of several research. Recently, we have proposed an approach which jointly exploits the information of the vocal tract and the glottis source. The approach synchronously takes into account the correlation between the two sources of information. The proposed theoretical model which consists of using a joint law is presented in this work. Some restrictions and simplifications were taken into account to show the significance of this approach in practical way. The fundamental frequency and the MFCC coefficients (Mel Frequency Cepstrum Coefficients) were used to represent the information of the source and the vocal tract, respectively. The probability density of the source, in particular, was considered to obey a uniform law. Tests were carried out with only the women speaker coming from de speech telephony database (SPIDRE) recorded from various hand set telephones. In this article, modelling the source information is proposed by using a Gaussian Mixture Model (GMM) rather than the uniform probabilistic model. Tests are extended to all speakers of the SPIDRE database. In this respect, four systems were proposed and compared. The first is a baseline system based on the MFCC and does not use any information from the source. The second examine only the voiced segments of the vocal signal. The last two relate to the suggested approaches according to the two techniques. The source information is supposed to follow a normal distribution in one technique and a logNormal distribution in the other. With the proposed approach, the profit in performance increases by 10,5% for the women, 7% for the men and 8% for all speakers.

1. INTRODUCTION

The most popular approach used to recognize the speaker identity is based on the use of the Mel Frequency Cepstral Coefficients (MFCC) as the parameters, and the Gaussians Mixtures Models (GMM) for the classification task [8]. The MFCC coefficients are supposed to extract and convoy the vocal track contribution. This approach worked well when the speech is recorded in clean conditions. However, the performance dropped considerably when speech was recoded in hostilitiy environments [9].

Recently, various approaches have been proposed to improve the performance of speaker identification systems. The main objectives be reached consist to be robust to: effect of the noise, channel distortions, handset variability and the number of speakers.

This work was funded by NSERC, Communications Security Establishment and the FUQAC. Many thanks to Karl Boutin for his support.

Particulary, several works were oriented to incorporate or combine in different way the source and vocal track information. The prosodic features are, usually, used as parameters for the glottic source. These are known to carry specific speaker information such as the melody, intonation and loudness. Generally, the melody and intonation are parameterized by the pitch (fundamental frequency) such averaged pitch, pitch contours, pitch jitters and location [1] [2] [6] [7]. On the other hand, the loudness is parameterized by the short-time spectral energy. Speaker recognition systems exclusively based on pitch do well when the number of speakers is small. However, performance decreases significantly when the number of speakers increases.

Indeed, the major problem is knowing how to incorporate and model glottis source and vocal tract information. Various techniques have been investigated for handling the unvoiced segments [11]. Shao and al. [12] have proposed an integrated pitch and MFCC extraction for speech recognition and reconstruction by using a auditory model. Zilca and al. [15] have presented a pseudo pitch synchronous algorithm for speaker recognition applications. Arcienega and Drygajlo [3] have presented a statical approach using pitch dependent GMMs. This approach purposes to model simultaneously the statical distribution of the short-term acoustic vectors and long-term prosodic features. In spite of the good results, the experiment is devoted only to clean speech. Most of the models reported in the literature assume the independence of the glottis and the vocal tract. Ezzaidi and al. [5] have proposed a statical model that takes into account the correlation between the glottis and the vocal tract. They used a joint probability function with many restrictions and hypotheses account for the correlation between source and vocal tract. Particulary, they used a uniform probabilistic model of the glottic source and GMM models for the MFCC coefficients.

In this paper, some modifications have been added to the model presented in [5]. A pitch is considered not locally uniform and a Gaussian Mixture model is proposed, tested and evaluated. A summary of the proposed model is described in the next section. The speech analysis, parameters estimation and database are described in section 3 , 4 and 5. The baseline and the proposed systems are described in section 6. Sections 7 and 8 present the results, discussion and conclusion.

2. Proposed model

Principles of the model suggested in a previous work [4] are outline here. The pitch ($\widehat{X}$) and vocal tract ($\widehat{Y}$) features are two discritezed random processes, and f is the joint probability of

the two discretized random processes $\widehat{X}$ and $\widehat{Y}$, so:

$$f(x_i, \overrightarrow{y_j}) = P(\widehat{X} = x_i, \widehat{Y} = \overrightarrow{y_j}) \text{ with} \quad (1)$$

$$0 \leq f(x_i, \overrightarrow{y_j}) \leq 1 \text{ and } \sum_{i=1}^{i=n} \sum_{j=1}^{j=m} f(x_i, \overrightarrow{y_j}) = 1. \quad (2)$$

The respective marginal probability functions are:

$$f(x_i) = \sum_{j=1}^{m} f(x_i, \overrightarrow{y_j}) \text{ and } f(y_j) = \sum_{i=1}^{n} f(x_i, \overrightarrow{y_j}). \quad (3)$$

Each speaker s is supposed to be defined by its probability function, as:

$$f_s(x_i, \overrightarrow{y_j}) = P_s(\widehat{X} = x_i, \widehat{Y} = \overrightarrow{y_j}). \quad (4)$$

It was observed that:

$$f_s(x_i, \overrightarrow{y_j}) \quad = \quad f_s(\overrightarrow{y_j}/x_i) f_s(x_i). \quad (5)$$

$f_s(x_i)$ is *a priori* probability of a pitch frequency equal to x_i and $f_s(\overrightarrow{y_j}/x_i)$ is the *conditional* probability of observing a feature vector equal to $\overrightarrow{y_j}$, given that the pitch frequency is x_i.

In the previous work, we have focused only on a local estimation and integration of conditional probability $f_s(\overrightarrow{y_j}/x_i)$. The factor $f_s(x_i)$, a prior probability in equation 5, was supposed locally uniform. The experiments are reported on the whole speakers Spidre database including men and women.

In this work, we focus on the conjoint integration of the $f_s(\overrightarrow{y_j}/x_i)$ and $f_s(x_i)$ in speaker recognition systems. A Gaussian Mixture Model was used for prior probability $f_s(x_i)$ and the conditional probability $f_s(\overrightarrow{y_j}/x_i)$.

2.1. Feature vector distributions based on pitch knowledge

Let us define I_k, $k = 1, \ldots, N$ as sub–intervals of the pitch set $\{x_1, x_2, .., x_n\}$. We recall that $x_1 = 66\ Hz$, $x_n = 660\ Hz$ and N is the number of intervals with $I_1 \cup \ldots \cup I_N = \{x_1, x_2, .., x_n\}$. Each subspace H_k in the space(x,$\vec{y}$) is associated with a pitch interval I_k. For each H_k, we suppose that the probability function $f_s(\overrightarrow{y_j}/x_i)$ is stationary and pitch dependent inside the interval I_k.

Theoretically, a number of models $f_s(\overrightarrow{y_j}/I_k) = \lambda_{s,k}$ would be equal to n and a number of models $f_s(\overrightarrow{x_i}) = \acute{\lambda}_{s,k}$ would also be equal to n. By subdividing the space into N subspaces, the number is reduced to N. The interval length of I_k is based on the shape of the pitch histogram.

3. Speech analysis

Mel Cepstrum Coefficients derived from a bank of filters (MFCC) are used as features to characterize the vocal tract information for the speaker identity. Coefficients of c_1 to c_{12} are used. The speech is first preemphasized (0.97); then, a sliding Hamming window with a length of 32 ms and a shift of 10 ms is positioned on the signal.

Cepstral mean normalization and liftering are also performed. Delta and delta-delta MFCC are not used, as the comparison between the systems would be biased. In fact, adjacent segments can have different pitch values belonging to different sub–intervals I_k.

The pitch is used as prosodic features to characterize the source contribution (glottic). Used in this work is the pitch tracker proposed by Rouat and al [10]. It is based on the computation of autocorrelation function estimated from each cochlear filters bancs. Particulary, the pitch tracker was proposed for the speech recorded from the telephony support, which is the case of the database used in this work. The pitch is estimated every 10 ms. A median filter is then used to smooth the pitch estimation over a window duration of 70 ms, in order to be less affected by the doubling/halving error pitch estimation.

4. Pattern recognition

4.0.1. Parametric model

A Gaussian Mixture Model (GMM) [8] is used as a parametric model. Each speaker is characterized by 2xN models, where N correspond to N-pitch intervals I_k. Precisely, N models are trained on a l-dimensional vector estimating the vocal tract contribution. Each model uses weighted sum of 32 ($M = 32$) Gaussians. The last N models are trained on 1-dimensional vector to estimate the pitch source information. Here, Each model uses a weighted sum of 4 ($M = 4$) Gaussians. This choice was justified by the histograms analysis. Thus each specific speaker s is characterize by two GMM models ($\lambda_{s,k}$ and $\acute{\lambda_{s,k}}$) for each pitch interval I_k, for the MFCC coefficient and the pitch information, respectively.

Let us define $p(\overrightarrow{y}/\lambda_{s,k})$, the Gaussian mixture density associated with the probability function $f_s(\overrightarrow{y_j}/I_k)$ for speaker s, as

$$p(\overrightarrow{y}/\lambda_{s,k}) = \sum_{i=1}^{M} w_{i,k} b_{i,k}(\overrightarrow{y}) \quad (6)$$

with

$b_{i,k}(\overrightarrow{y}) = \frac{1}{(2\pi)^{l/2}|\Sigma_{i,k}|^{1/2}} exp\{-\frac{1}{2}(\overrightarrow{y} - \overrightarrow{\mu_{i,k}})' \Sigma_{i,k}^{-1}(\overrightarrow{y} - \overrightarrow{\mu_{i,k}})\}$.

M is the GMM order, $\overrightarrow{y}$ is the l-dimensional vector estimating. The i-th Gaussian density is noted as $b_{i,k}$ with mean $\mu_{i,k}$ and covariance matrix $\Sigma_{i,k}$ and $w_{i,k}$ are the mixture weights. $b_{i,k}$, $\mu_{i,k}$, $\Sigma_{i,k}$ and $w_{i,k}$ are defined for pitch interval I_k and for speaker s.

In the context of the framework, the same model was proposed for the pitch information. The number of mixture of Gaussians was fixed to $M = 4$.

4.1. Recognition criterion

We define T as the test length over which the recognition is performed. A frame-by-frame estimation of log–likelihood for each speaker s and pitch interval I_k is first performed.

Each frame (32 ms length) is shifted by 10 ms. Then, the maximum log–likehood for each speaker is estimated over T. When the test sentence is longer than T, the score average over a number of length T is computed according to this equation:

$$S_{T,s} = \frac{nb.\ of\ seg.\ correctly\ tested\ for\ T\ duration}{total\ nb.\ of\ seg.\ tested\ for\ T\ duration} \quad (7)$$

The final identification score is obtained by averaging over the number of speakers N_s, as:

$$Score = \frac{\sum_{i=1}^{Ns} S_{T,i}}{Ns} \quad (8)$$

5. Speech database

A SPIDRE–Swichboard Corpus is used. It is comprised of a 45 speaker database, including all men and women. Each speaker has 4 conversations originating from 3 different handsets. The

training data contains 3 conversations, with 2 conversations coming from the same handset. The last conversation, using the third handset (different from the others), is presented as the test data. This combination is referred to as the *mismatched condition*.

6. Strategies

6.1. The baseline strategy

The baseline strategy uses both the voiced and unvoiced segments. The suppression of silence was carried out based on the energy evolution and the comparison with fixed thresholds. One model, $\lambda_{s,b}$, for each speaker is generated for the baseline system.

6.2. Recognition based on voiced speech segments

A module that estimates the pitch and selects the voiced segments is included. A pitch tracker and a voiced-unvoiced detection system [10] in conjunction with the SID system analysis module are used as well. In this case, silence and unvoiced segments are automatically rejected. During training for each pitch period, we centered a 32 ms duration window and extracted the MFCC coefficients.

One model, $\lambda_{s,v}$, is generated for recognition on the voiced speech system.

6.3. Recognition based on the estimated a posteriori probabilites

For the third and the fourth proposed systems, four pitch intervals I_1,...,I_4 are created according to the pitch frequency histogram. More than 90% of the pitch frequencies belong to the interval of [150Hz,220Hz] for women speakers and to the interval of [90Hz,150Hz] for men speakers. The pitch frequencies are distributed over 4 intervals as follows:

- Women intervals :
  - $I_1 = [150, 180]$;
  - $I_2 = [170, 200]$;
  - $I_3 = [190, 220]$;
  - $I_4 = [63, 150] \cup [220, 600]$.
- Men intervals:
  - $I_1 = [90, 120]$;
  - $I_2 = [110, 130]$;
  - $I_3 = [120, 150]$;
  - $I_4 = [63, 90] \cup [150, 600]$.

The choice of four intervals is a trade off between fine pitch intervals and sufficient training size of the models. During training and for each interval I_k, the MFCC vectors and pitch are used to generate a parametric model for each speaker. Therefore each speaker is characterized by 2x4 models. With the aim of overcoming the pitch estimation errors, we choose an overlap of 10 Hz between the intervals. Thus, the MFCC vectors from speech whose fundamental frequency belongs to two adjacent intervals (I_k, I_{k+1}), associated to subspaces H_k and H_{k+1}, respectively are during the testing session, evaluated over these two subspaces and we keep the best score.

The fourth model $\lambda^p_{s,k}$ is generated for recognition, taking into account the conditional probability of voiced speech according to the pitch. The pitch is distributed as a Gaussians Mixture. The difference between the two proposed systems simply consists in training the third system model with the logarithm of the pitch instead of the estimated pitch value. The third technique is inspired from the work of S*ö*nmez and al.[13]. Using a simple correlation model, they showed that the pitch has a lognormal distribution. Therefore, in the third system the training is done with log(pitch). On the other hand, in the fourth system the training is directly carried out data pitch without any postprocessing.

7. Results and discussion

Table 1: scores of all speakers

times	100 ms	500 ms	1 s	2 s	3 s
baseline	29%	58%	71%	81%	85%
voiced	+1.48	+1.38	+1.72	+3.11	+3.17
LogFoMod	+9.32	+6.17	+3.04	+2.05	+0.14
FoMod	+9.30	+6.10	+3.06	+2.34	+0.28

Table 2: scores of women

times	100 ms	500 ms	1 s	2 s	3 s
baseline	26%	56%	68%	79%	84%
voiced	+0.91	+0.69	+1.84	+3.70	+3.23
FoLogMod	+11.84	+9.42	+6.82	+4.74	+0.76
FoMod	+11.87	+9.09	+6.63	+4.83	+0.76

Table 3: scores of men

times	100 ms	500 ms	1 s	2 s	3 s
baseline	31%	61%	73%	82%	86%
voiced	+5.71	+4.58	+3.39	+3.61	+3.73
FoLogMod	+10.12	+4.80	+0.55	+0.07	-0.34
FoMod	+10.08	+4.87	+0.70	+0.49	-0.09

Tables 1, 2 and 3 report the identification results observed with the four techniques: 1) Baseline (voiced and unvoiced segments), 2) Voiced (only voiced segments), 3) Voiced segments with partition of space into H_1 to H_4. The pitch is supposed to distribute according to the lognormal model (noted LogFoMod) and 4) Voiced segments with partition of space into H_1 to H_4. The pitch is supposed to distribute according to normal model (noted FoMod).

The first column gives the value of T, that is, the duration of maximum log–likehood estimation.

Compared to the baseline system, the sign '+' indicates a profit in performance and '-' indicates a loss in performance.

The results show that the baseline system yields the lowest identification rates. All the proposed systems yield a profit in score for all durations and in the reported experiments. The profit in score decreased as the time duration test increased. Particulary, the proposed system yields a profit in score of almost 10% for women speakers, 7% for men speakers and 8% for all speakers, when the test duration is less than 500 ms. Consequently, we can deduce that the strategy suggested is interesting for systems which make decisions over short durations. It should be emphasized here that almost all systems converge

towards the same weak scores when the test duration is relatively large. This weakening in profit can be explained by the non-standardization and the different space-dimensions of MFCC coefficients and the pitch. In particular, the pitch variation which is considered weak compared to MFCC implies that de pitch probability density is narrow resulting in biased decisions when time duration increases. Consequently, it will be judicious to balance the probability densities in order to obtain homogenized scores.

Generally the best score is observed with men speakers and the weaker score is observed with women speakers. Comparing the last two techniques, we found that they yielded similar scores. Therefore, modelling the pitch by a lognormal distribution can be considered unsuitable and inappropriate for the speech with telephone quality.

For the results presented, a T of 1 second is equivalent to 100 MFCC vectors and is independent of the technique. The weak performance of the baseline system might be partially due to the smaller number of voiced frames in a fixed T.

As shown in tables 1, 2 and 3, the proposed models improve the baseline system. When the dependence of source and vocal tract is taken into account, the best results are observed for durations, T, lower than 500 ms.

In several cases, the pitch is not well estimated and affects the performance. If these errors are corrected, we can possibly contribute to better training and evaluation.

8. Conclusion

Motivated by the fact, that the speaker intervariability is more apparent on the basis of the pitch histograms rather that the spatial distribution for formants, an approach that preserves the dependence between the vocal source and the vocal tract has been proposed. Experiments that integrate the a-posteriori probability of observing a MFCC vector given the knowledge of the pitch frequency have been reported. The MFCC and pitch parameters are modelled respectively by 32 and 4 mixtures of Gaussians. They are compared with a baseline system operating on all voiced and unvoiced speech segments and with a second system that operates on voiced speech segments only. Closed set Speaker Identification experiments were performed on the SPIDRE corpus which comprises highly confusable female speakers.

Systems based on voiced segments yield good scores. However, when the dependence of the source and vocal tract is taken into account, the best results are observed for durations T lower than 500 ms (10% for women speakers, 7% for men speakers,8% for all speakers).

Despite the limited improvement in performance, it appears that the approach is promising. In fact, many restrictive hypotheses have again been made to set up the experiments as: the pitch tracker has been supposed to be reliable; a sufficient training data is assumed for subspace decomposition; a direct dependence of MFCC and pitch in each subspace is assumed.

We therefore suggest, as future work, to optimize the number/width of the pitch intervals (I_k) and to introduce weighting between a priori probability distribution of the pitch ($f_s(x_i)$) and conditional probability ($f_s(\overrightarrow{y}_j/x_i)$) in accordance with equation 5. A modified version of the proposed method should be investigated, in order to keep and to exploit conjointly the unvoiced and the voiced speech segment.

9. References

[1] Atal B. S., Automatic speaker recognition based on pitch contours, In *The Journal of the Acoustical Society of America*, pp. 1687-1697, Vol. 52, 1972.

[2] Atal B. S., Automatic recognition of speakers from their voices, In *Proc. IEEE*, pp. 460-475, Vol. 64, 1976.

[3] Arcienega M. and Drygajlo A., Pitch-dependent GMMs for Text-Independent Speaker Recognition Systems, In *Eurospeech'01*, 2001, Scandinavia.

[4] Hassan Ezzaidi and Jean Rouat, Comparison of MFCC and pitch synchronous AM, FM parameters for speaker identification, In *ICSLP*, October 2000.

[5] Hassan Ezzaidi, Jean Rouat and DouglasOShaughnessy, Towards combining pitch and MFCC for speaker identification systems, In *proceedings of Eurospeech* , September 2001.

[6] Jankowski C.R, Quatieri T.F. and Reynolds D.A., Measuring fine structure in speech: Application to speaker identification, In *IEEE ICASSP*, pp. 325-328, 1995.

[7] Plumpe M.D, Quatieri T.F., and Reynolds D.A, Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification, In *IEEE Transactions on Speech and Audio*, 1999.

[8] Reynolds Douglas A., A gaussian mixture modeling approach to text independent speaker identification, *Thesis, Georgia Institute of Technology*, August 1992.

[9] Reynolds Douglas A., The effects of handset variability on speaker recognition performance: Experiments on the Switchboard corpus, In *IEEE ICASSP*, 1996.

[10] J. Rouat and Yong Chun Liu and D. Morissette, A pitch determination and voiced/unvoiced decision algorithm for noisy speech, In *Speech Communication*, Vol. 21, pp. 191–207, 1997,

[11] K. Tokuda, T. Masuko and T. Kobayashi, Hidden Markov Models based on multi-space probability distribution for pitch pattern modeling,In *IEEE International Conference On acoustics*, Speech and Signal Processing, Vol. 1, pp. 229–232, March, 1999.

[12] X Shao, B. Milner and S. Cox Integrated Pitch and MFCC Extraction for Speech Reconstruction and Speech Recognition Applications, In *Eurospeech'03, Geneva*, pp. 1725-1728, 2003.

[13] Kemal S*ö*nmez, Larry Heck, Mitchel Weintraub and Elisabeth Shriberg, A lognormal tied mixture model of pitch for prosody-based speaker recognition, *In Proc. of Eurospeech*, pp. 1391-1394, 1997.

[14] Kemal S*ö*nmez, Elisabeth Shriberg, Laryy Heck and Mitchel Weintraub, Modeling dynamic prosodic variation for speaker verification, In *Proc. of International Conference on Spoken Language Processing*, pp. 3189-3192, 1998.

[15] R. D. Zilca, J. Navratil and G. N. Ramaswamy, "SynPitch": A pseudo Pitch Synchrounous Algorithm For Speaker Recognition, In *Eurospeech'03, Geneva*, pp. 2649-2652, 2003.