

TOWARDS COMBINING PITCH AND MFCC FOR SPEAKER IDENTIFICATION SYSTEMS

Hassan Ezzaidi, Jean Rouat and Douglas O'Shaughnessy⁺

Ermetis, Université du Québec à Chicoutimi, Chicoutimi, Québec, Canada, G7H 2B1

⁺INRS-Télécommunications, Université du Québec

900 de la Gauchetiere west, Box 644, Montreal, Québec, Canada, H5A 1C6

ABSTRACT

Usually, speaker recognition systems do not take into account the dependence between the vocal source and the vocal tract. A feasibility study that retains this dependence is presented here. A model of joint probability functions of the pitch and the feature vectors is proposed. Three strategies are designed and compared for all female speakers taken from the SPIDRE corpus. The first operates on all voiced and unvoiced speech segments (baseline strategy). The second strategy considers only the voiced speech segments and the last includes the pitch information along with the standard MFCC. We use two pattern recognizers: LVQ-SLP and GMM. In all cases, we observe an increase of the identification rates and more specifically when using a time duration of 500ms (6% higher).

1. INTRODUCTION

The vibration frequency of the vocal folds is known to be an important feature to characterize speech and has been found effective for automatic speech and speaker recognition [1] [6]. An important characteristic of pitch is its robustness to noise and channel distortions. Many parametrizations of pitch such as pitch value, averaged pitch, pitch contour, pitch jitter and location [1] [4] have been proposed for speaker verification or identification. Speaker recognition systems exclusively based on pitch do well when the number of speakers [1] is small. However, performance decreases significantly when the number of speakers increases, but pitch information can be reliably used to distinguish the sex of speakers [5]. In spite of the weak contribution of pitch to contemporary speaker identification, it remains true that the mechanisms involved in speech production are complex, and imply dependence of articulators and vocal folds, which can be useful for speaker verification or identification.

The most popular way used to model pitch is by a Gaussian density or a mixture of Gaussians. Statistical independence of the glottis and the vocal tract is assumed by these models. In this paper, we propose to take into account the correlation between the glottis and the vocal tract. We study the influence of this dependence in the context of a text-independent Speaker Identification System (SIS). We use a joint probability function to take into account the correlation between source and vocal tract. The proposed approach consists of generating models of the feature vectors for each pitch range.

hezzaidi@uqac.quebec.ca, jrouat@uqac.quebec.ca, dougo@inrs-telecom.quebec.ca

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

The next section describes the motivation for this research. The baseline and the proposed systems are described in sections 5 and 7. Sections 8 and 9 present the results, discussion and conclusion.

2. MOTIVATION

Most systems use parameters that encode vocal tract features, but contributions of the glottis are largely ignored. Even if MFCC are theoretically known to deconvolve the source and the vocal tract, in practice, cepstrum coefficients are affected by high pitched voices (women and infants). One can illustrate the role of pitch when dependence of the source and the vocal tract are maintained. Figure 1 exhibits four spectrograms and pitch histograms, each column corresponds to a different male speaker, obtained from the YOHO database. All speakers pronounced the same digit utterance 'twenty six'. The pitch range is divided into 56 linearly bins of 10 Hz width. The spectrograms show a significant similarity of formant distributions between speakers. The spatial distribution of formants depends on the interspeaker variability as described in [2]. However, the pitch histograms are different and vary from one speaker to another for the same context. If one compares the histograms by taking into account their frequency amplitude and width, it is observed that speaker 2 from the second column and speaker 4 from the fourth column do have a similar pitch distribution. On the other hand, speakers 1 and 3 are characterized by

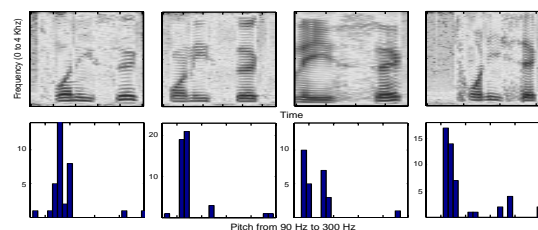


Figure 1: For each of four male speakers: Pitch histograms and spectrograms for the same English digit utterance '26'.

dissimilar pitch histograms. Consequently, if one takes into consideration the pitch information, the interspeaker variability can be restricted to speakers with similar pitch distributions, and the other speakers will be considered as belonging to other clusters. Speakers with similar pitch will be recognized based on the spectral characteristics.

In summary, pitch and vocal tract features can be jointly exploited in order to establish probability models of feature vectors

assuming the a priori knowledge of the pitch distribution.

3. PROPOSED MODEL

3.1. Theoretical framework

We suppose that pitch and vocal tract features are two random process respectively denoted as $X(t)$ and $Y(t)$. Let's write $\widehat{X}(n)$, the estimated discretized pitch frequency at time $n\Delta t$ and $\widehat{Y}(n)$ the estimated discretized vocal tract feature vector at time $n\Delta t$. $\widehat{Y}(n)$ is an l -dimensional vector. In practice $\widehat{Y}(n)$ is an *LPC* or *MFCC* vector estimated from a centered signal window at time $n\Delta t$. For each process $\widehat{X}(n)$ and $\widehat{Y}(n)$, we assume time independence of their respective realizations. As a consequence, $\widehat{X}(n+1)$ is independent of the realization of $\widehat{X}(n)$. The same restriction applies to $\widehat{Y}(n)$. In the following, we drop the time and consider the simultaneous realization of $\widehat{X}(n)$ and $\widehat{Y}(n)$ as being time independent. The crosscorrelation between $\widehat{X}(n)$ and $\widehat{Y}(n)$ is still preserved.

Let us write $\{x_1, x_2, \dots, x_n\}$, the increasing sequence of realizations of \widehat{X} , with $x_i \in [60 \text{ Hz}, 660 \text{ Hz}]$. We suppose that the set of realizations of \widehat{Y} is finite (by using vector quantization for example) and equal to $\{\vec{y}_1^l, \vec{y}_2^l, \dots, \vec{y}_m^l\}$, with $\vec{y}_i^l \in R^l$. Let f to be the joint probability of \widehat{X} and \widehat{Y} .

$$f(x_i, \vec{y}_j^l) = P(\widehat{X} = x_i, \widehat{Y} = \vec{y}_j^l) \text{ with} \quad (1)$$

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

The respective marginal probability functions are:

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

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

$$f_s(x_i, \vec{y}_j^l) = P_s(\widehat{X} = x_i, \widehat{Y} = \vec{y}_j^l). \quad (4)$$

We observe that

$$f_s(x_i, \vec{y}_j^l) = f_s(\vec{y}_j^l/x_i) f_s(x_i) \quad (5)$$

$f_s(x_i)$ is the *a priori* probability of a pitch frequency to be equal to x_i and $f_s(\vec{y}_j^l/x_i)$ is the *a posteriori* probability of observing a feature vector to be equal to \vec{y}_j^l given the knowledge of the pitch frequency x_i . The estimation of the a priori probability of the pitch frequency is relatively straightforward while the estimation of $f_s(\vec{y}_j^l/x_i)$ can be long and tedious.

3.2. Feature vector distributions based on pitch knowledge

In the present work we focus on the estimation and integration of the posteriori probability, $f_s(\vec{y}_j^l/x_i)$, in speaker recognition systems. The consideration of the factor $f_s(x_i)$ from equation 5 is left as a future work.

We propose to subdivide the space (x, \vec{y}) into subspaces H_k where $f_s(\vec{y}_j^l/x_i)$ is supposed to be locally independent of the pitch value. Let us define I_k , $k = 1, \dots, N$ as sub-intervals of the pitch set $\{x_1, x_2, \dots, x_n\}$. We recall that $x_1 = 66 \text{ Hz}$ and $x_n =$

660 Hz , N is the number of intervals with $I_1 \cup \dots \cup I_N = \{x_1, x_2, \dots, x_n\}$. Each subspace H_k is associated to a pitch interval I_k . For each H_k , we suppose that the probability function is stationary and independent of the pitch inside the interval I_k , that is,

$$f_s(\vec{y}_j^l/x_i) \simeq f_s(\vec{y}_j^l/I_k). \quad (6)$$

Theoretically, the number of models $f_s(\vec{y}_j^l/x_i)$ would be equal

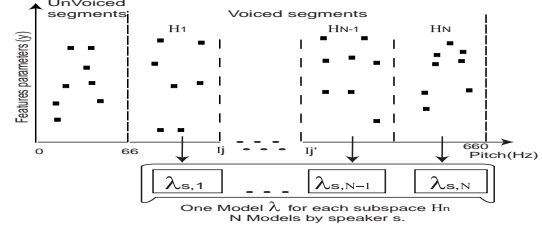


Figure 2: Proposed approach for generating sub-models.

to n . By subdividing the space into N subspaces, we reduce that number to N . Figure 2 illustrates the notion of subspaces and models of probability functions $f_s(\vec{y}_j^l/I_k)$. The interval length of I_k is based on the shape of the pitch histogram (section 7.3).

4. SPEECH ANALYSIS

Mel Cepstrum Coefficients derived from a bank of filters (MFCC) are used as features to characterize the identity of speakers. We use coefficients c_1 to c_{12} . 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 .

5. PATTERN RECOGNITION

5.1. Framework

Two pattern recognizers are used for the experimental task: one parametric and one non-parametric.

5.1.1. Parametric model

We use a Gaussian Mixture Model (GMM) [7] with 32 ($M = 32$) weighted sums of Gaussians. Each GMM is defined for a specific speaker s and pitch interval I_k . Let us define $p(\vec{y}/\lambda_{s,k})$, the Gaussian mixture density associated to the probability function $f_s(\vec{y}_j^l/I_k)$ for speaker s , as

$$p(\vec{y}/\lambda_{s,k}) = \sum_{i=1}^M w_{i,k} b_{i,k}(\vec{y}) \quad (7)$$

with

$$b_{i,k}(\vec{y}) = \frac{1}{(2\pi)^{l/2} |\Sigma_{i,k}|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{y} - \vec{\mu}_{i,k})' \Sigma_{i,k}^{-1} (\vec{y} - \vec{\mu}_{i,k})\right\}.$$

M is the GMM order, \vec{y} is the l -dimensional vector estimating the vocal tract contribution (MFCC vector), $b_{i,k}$ is the i -th Gaussian density 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 . Each speaker is characterized by N models $\lambda_{s,k}$ corresponding to N pitch intervals I_k .

5.1.2. Non-parametric model

We use the hybrid LVQ-SLP network as proposed by He *et al.* [3]. Each speaker s , with a pitch belonging to I_k , is characterized by a codebook $C_{s,k}$. The codebook size is the same for all speakers. We performed experiments with codebook sizes of 512 for each speaker.

5.2. Recognition

5.2.1. Parametric model

We define T as being 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-likelihood for each speaker is estimated over T . When the test sentence is longer than T , the average of the score over the number of segments with a length of T is computed according to equation 8.

$$S_T^s = \frac{\text{nb. of seg. correctly tested for } T \text{ duration}}{\text{total nb. of seg. tested for } T \text{ duration}} \quad (8)$$

The final identification score (equation 9) is obtained by averaging over the number of speakers N_s :

$$\text{Score} = \frac{\sum_{i=1}^{N_s} S_T^s}{N_s} \quad (9)$$

5.2.2. Non-parametric model

For each frame, the feature vector is classified by using the Nearest Neighbor criteria. A speaker is recognized if, for the entire test conversation, it is selected more frequently than the other speakers.

6. SPEECH DATABASE

A subset of the SPIDRE-Swichboard Corpus is used and comprises eighteen (18) female speakers of the database. 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*. The *matched condition* refers to situations where training and testing data are recorded from the same handset.

7. STRATEGIES

7.1. 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 comparison with fixed thresholds.

7.2. Recognition based on voiced speech segments

We include a module that estimates the pitch and selects the voiced segments. We use a pitch tracker and a voiced-unvoiced detection system [8] in conjunction with the SID system analysis module. In this case, silence and unvoiced segments are automatically rejected. During training and for each pitch period, we centered a 32 ms duration window and extracted the MFCC coefficients.

7.3. Recognition based on the estimated a posteriori probabilities

For the third strategy, four pitch intervals I_1, \dots, I_4 are created according to the pitch frequency histogram. More than 90% of the pitch frequencies belong to the interval [150Hz, 220Hz]. We distributed the pitch frequencies over 4 intervals $I_1=[150,180]$, $I_2=[170,200]$, $I_3=[190,220]$ and $I_4=[66,150] \cup [220,660]$. 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 are used to generate model parameters for each speaker. Therefore each speaker is characterized by 4 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}), will be used to train two models, respectively, associated to subspaces H_k and H_{k+1} . Then, during the testing session, the evaluation is carried out over these two subspaces and we keep the best score.

In the case of LVQ-SLP, the codebook generation is made according to two procedures. One attributes the same codebook size to each subspace, and the other distributes the number of prototypes per codebook according to the number of events in each subspace.

In the case of the GMM models, one model λ_s is generated for the baseline system, one model is also used for recognition on voiced speech and four models $\lambda_{s,k}$ are generated for the recognition taking into account the a posteriori probabilities of voiced speech according to the pitch.

8. RESULTS AND DISCUSSION

8.1. Evaluation with a LVQ-SLP model

LVQ-SLP results for 18 women of the SPIDRE database are reported in tables 1 and 2.

Table 1: LVQ-SLP: Identification rate increases for 18 female speakers with fixed codebook sizes. Baseline system: 55%.

| | Voiced (512) | H_1 (128) | H_2 (128) | H_3 (128) | H_4 (128) |
|------------|-----------------|----------------|----------------|----------------|----------------|
| Matched | 6 | 14 | 10 | 1 | 0 |
| Mismatched | 3 | 4 | 4 | 5 | 2 |

When the unvoiced segments are not taken into account (Voiced column), the identification rate increases to 61% (6% more in table 1) for matched handsets and to 58% (3%) for mismatched handsets. When pitch is taken into account (columns H_k in tables 1 and 2), the increase is almost the double in H_1 and H_2 and weaker in H_3 and H_4 . H_1 and H_2 are the subspaces with the greatest number of events and H_3 and H_4 with the smallest number of events. When the number of prototypes per codebook is

Table 2: LVQ-SLP: Identification rate increases for 18 female speakers with codebook sizes proportional to the number of events in each subspace. Baseline system: 55%.

| | H_1 | H_2 | H_3 | H_4 |
|------------|-------|-------|-------|-------|
| Matched | 14 | 3 | 0 | -1 |
| Mismatched | 4 | 5 | 6 | 1 |

proportional to the number of events, performance falls in H_2 and H_4 and remains constant in H_1 .

When recognition rates are weighted according to the number of events per subspace, we obtain an averaged increase of 8% in matched conditions, and 4% in mismatched situations. It is observed that the identification results are sensitive to several factors: 1) codebook sizes, 2) training techniques, 3) scores combination, and 4) pitch estimation. The best increase in performance is observed for subspaces with the greatest number of events.

8.2. Evaluation with a GMM model

Table 3 reports the identification results observed with the three strategies: 1) Baseline (voiced and unvoiced segments), 2) Voiced (only voiced segments) and 3) Voiced segments with partition of space into H_1 to H_4 . The first column gives the value of T , that is, the duration of maximum log-likelihood estimation. The baseline

Table 3: GMM: Mismatched identification rates for 18 female speakers.

| Time(seconds) | Baseline(%) | Voiced(%) | Voiced & pitch(%) |
|---------------|-------------|-----------|-------------------|
| 0.1 | 36.8 | 37.7 | 40.5 |
| 0.5 | 63.4 | 66.7 | 69.4 |
| 1 | 75.4 | 79.9 | 80.8 |
| 2 | 84.2 | 87.9 | 88.0 |
| 3 | 88.0 | 90.6 | 90.5 |
| 4 | 90.0 | 93.9 | 93.3 |
| 5 | 91.4 | 95.4 | 94.7 |
| 6 | 92.7 | 95.3 | 95.2 |

strategy yields the lowest identification rates. When voiced segments are used, the best increase is 4.5% and the weakest is 1%. When a preliminary subdivision based on pitch is performed, the greatest increase is 6% and the weakest is 2.5%.

When the test duration is greater than 2 seconds, strategies based on voiced segments yield similar results. When T is less than 2 seconds, the best results are observed with the strategy that takes into account the posterior probability (subdivision into subspaces). The increase is on the order of 2%.

8.3. Discussion

Identification rates of LVQ-SLP and GMM are not strictly comparable, as the recognition criteria is different. The comparison of Tables 1 and 2 suggests that the a priori probability $f_s(x_i)$ should be taken into account. In Table 3, a T of 1 second is equivalent to 100 MFCC vectors and is independent of the strategy. The weaker performance of the baseline system might be partially due to the smaller number of voiced frames in a fixed T . In several cases, the pitch is not well estimated and affects the performance. If these errors are corrected, we can possibly achieve a better training and evaluation. The voiced pitch strategy requires more calculation and is more sensitive to errors especially during training.

9. CONCLUSION

A new 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. 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 a subset of the SPIDRE corpus that comprises highly confusable female speakers. Systems based on voiced segments yield the best scores.

When the dependence of the source and vocal tract is taken into account, the best results are observed for durations T lower than 3 seconds (up to 4.5% for $T = 500$ ms). For $T \geq 3$ seconds scores are 1% higher, in favour of the system based on voiced segments only.

Despite the small improvement in performance, it appears to us that the approach is promising. In fact, many restrictive hypotheses have been made to set up the experiments. The pitch tracker has been supposed to be reliable; sufficient training data for subspaces decomposition, local independence of MFCC in relation to pitch in a subspace (equation 6), and time independence of pitch and MFCC have been assumed.

We therefore suggest, as future work, to increase the size of the corpus for a better statistical convergence, to optimize the number and width of the pitch intervals (I_k) and to introduce weighting by the a priori probability distribution of the pitch ($f_s(x_i)$) in accordance with equation 5. We also suggest restricting the application to a text-dependent system, for which the variability of the parameters is usually smaller.

10. REFERENCES

- [1] B. S. Atal. Automatic recognition of speakers from their voices. In *Proc. IEEE 1976*, volume 64, pages 460–475, 1976.
- [2] G. R. Doddington. Speaker recognition-identification people by their voices. In *Proc. IEEE Vol.73, No 11*, number 11, pages 1651–1664, 1985.
- [3] J. He, L. Liu, and G. Palm. Speaker identification using hybrid lvq-slp networks. In *Proc. IEEE ICNN Vol.4*, pages 2051–2055, 1995.
- [4] C.R Jankowski Jr., T.F. Quatieri, and D.A. Reynolds. Measuring fine structure in speech: Application to speaker identification. In *IEEE-ICASSP*, pages 325–328, 1995.
- [5] J.Rouat, H. Ezzaidi, and M. Lapointe. Nouv. alg. d’extract. en vue de caract. le loc. Technical report, March 1999. Contrat W2213-9-2234/SL, 67 pages.
- [6] Douglas O’Shaughnessy and Hesham Tolba. Towards a robust/fast continuous speech recognition system using a voiced-unvoiced decision. pages 413–416, ICASSP, 1999.
- [7] Douglas A. Reynolds and Richard C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. 3(1):72–83, 1995.
- [8] J. Rouat, Y.C. Liu, and D. Morissette. A pitch determination and voiced/unvoiced decision algorithm for noisy speech. *Speech Communication*, 21:191–207, 1997.