Analysis of band structures for speaker-specific information in FM feature extraction

Tharmarajah Thiruvaran, Eliathamby Ambikairajah and Julien Epps

School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney NSW 2052 Australia.
National Information Communication Technology (NICTA),
Australian Technology Park, Eveleigh 1430, Australia

thiruvaran@student.unsw.edu.au, ambi@ee.unsw.edu.au, j.epps@unsw.edu.au

Abstract

Frequency modulation (FM) features are typically extracted using a filterbank, usually based on an auditory frequency scale, however there is psychophysical evidence to suggest that this scale may not be optimal for extracting speaker-specific information. In this paper, speaker-specific information in FM features is analyzed as a function of the filterbank structure at the feature, model and classification stages. Scatter matrix based separation measures at the feature level and Kullback-Leibler distance based measures at the model level are used to analyze the discriminative contributions of the different bands. Then a series of speaker recognition experiments are performed to study how each band of the FM feature contributes to speaker recognition. A new filter bank structure is proposed that attempts to maximize the speaker-specific information in the FM feature for telephone data. Finally, the distribution of speaker-specific information is analyzed for wideband speech.

1. Introduction

Perhaps in the majority of speech front-ends, the speech signal is decomposed into subbands, whose spacing and bandwidths are often motivated by auditory frequency scales. Despite the success of front-ends simulating the process of auditory perception, it has been shown that machine performance in speaker recognition tasks is very competitive with human performance, as far back as about a decade [1]. Here, although the machine did not exactly replicate human perception of speech, it is still competitive. This may be because the machine can be tailor-made specific to speaker recognition, which further implies that each part of the feature extraction process could be tested and tuned to speaker-specific information.

There are psychophysical explanations for the speaker-specific information being concentrated in particular frequency regions. The effect of the hypopharynx (laryngeal tube and the piriform fossa) was found to be relatively stable for vowel production but varied widely among different speakers [2]. This inter-speaker variation affects the frequency spectra above approximately 2.5 kHz. In addition, the higher formants such as F4 and F5 have been found to be rather stable for continuous speech [3]. F4 is determined almost entirely by the geometry of the laryngeal cavity, which has been suggested as an organ to transmit non-linguistic information. Further, the fundamental frequency in the range between 50 Hz to 400Hz depends on the stiffness and length of the vocal folds[4]. Apart from these psychophysical explanations, an experimental analysis was performed using MFCC features on a speaker recognition database with 18 kHz sampling frequency [4]. The authors experimentally found that the three frequency regions from 100 Hz to 300Hz, from 4 kHz to 5.5 kHz and from 6.5 kHz to 7.8 kHz contain higher speaker-specific information. Further, the frequency region from 500Hz to 3.5 kHz was found to have less speaker discriminative information. Unfortunately the bandwidth of the major speaker recognition databases, the NIST SRE databases, lies mainly within this region. This paper studies how speaker-specific information in frequency modulation features is distributed among different bands mainly on telephone bandwidth of NIST SRE database using FM feature not only at the feature level as in [4], but also at model and classification levels. FM is selected as each dimension of FM directly corresponds to each subband [5]. Psychophysical and experimental motivation for the use of FM in speaker recognition was introduced in [5-6].

FM can characterize the phase of the signal, and its importance has emerged recently as a complementary feature to amplitude-based features [5]. Nonlinearities in speech production, together with the evidence for modulation in speech production led to an AM-FM modulation model being proposed in [7]. Here, modulation around the vocal tract resonances is explained as “the air jets flowing through the vocal tract during speech production is highly unstable and oscillates between its walls, attaching or detaching itself, and thereby changing the effective cross-sectional areas and air masses, which affects the frequency of the cavity resonator” [7]. FM feature extraction involves subband filtering and usually auditory motivated Bark or Mel scale filters are used.

This paper analyses how speaker-specific information is distributed across different bands in FM feature extraction, with a view to reallocate filter bank centre frequencies and bandwidths specifically for speaker recognition in the telephone bandwidth. Further, this paper will assess whether the auditory motivated filter bank is a good choice for speaker recognition. For reasons of computational convenience, the relatively small NIST2001 database has been used extensively and then selected experiments are conducted on NIST 2006 male database. Finally this study is extended to wideband speech sampled at 44.1 kHz using the CHAINS corpus [8].

2. FM feature extraction

Based on the evidence for the existence of modulations in speech, the resonances are each modeled as AM-FM signals in, and summed, can be used to model the speech as in (1) [7].

\[ x(t) = \sum_{k=1}^{K} A_k(t) \cos \left( 2\pi f_k t + 2\pi \int q_k(r) d\tau + \theta_k \right) \]  

(1)
where $K$ is the total number of resonances, $A_d(t)$ is the time-varying amplitude component, $q(t)$ is the time-varying frequency component, $f_c$ is the center frequency of the resonance and $q_0$ is the initial phase. This model is the basis for FM extraction from speech.

In this work, frame averaged FM features are extracted over a 30 ms window length, mainly using the second-order all-pole method [5]. In this method, speech is initially filtered using Gabor bandpass filters and FM is extracted in each subband. Then each subband signal is modeled as a second order AR process, from which the pole frequency is estimated. From the pole frequency, which is taken as an estimate of the instantaneous frequency, the FM component is calculated by subtracting the center frequency of the subband. The dimension of the FM feature is 14, directly corresponding to 14 subbands [5].

3. Speaker specific information at the feature and model levels

Speaker-specific information of FM in each band is initially analyzed by studying which dimension of FM contributes more to speaker recognition. The variation of speaker separation across different bands is studied in feature space using scatter matrix based separation index ($J$) and in model space using the KL distance, calculated for all 174 speakers in NIST 2001 training data. A filter bank with 14 filters uniformly spaced across the telephone bandwidth (300Hz to 3700Hz) is used to decompose the speech. For this study uniform scale is preferred to the usual Bark scale filters in order to avoid the bias due to the increasing bandwidths of Bark scale. As the inter-dimensional correlation of FM features is negligible [9], dimensional independence is assumed in these analyses.

3.1. Scatter matrix based separation index ($J$)

The scatter matrix based separation index ($J$) measures how far the classes are separated. For a data matrix with $N$ columns $X_i$ ($i$=1 to $N$) of $d$ dimensional data of $K$ classes the separation index ($J$) is defined as [10]

$$J = tr(S_B^{-1}S_W)$$

(2)

where $S_B$ is the within-class scatter matrix and $S_W$ is the between-class scatter matrix defined in equations 3 and 4 respectively and $tr()$ is the trace operator.

$$S_B = \sum_{j=1}^{K} N_j (\bar{X}_j - \bar{m}_j)(\bar{X}_j - \bar{m}_j)^T$$

(3)

$$S_W = \sum_{j=1}^{K} N_j (\bar{m}_j - \bar{m})(\bar{m}_j - \bar{m})^T$$

(4)

where $\bar{m}_j$ is the mean of the class $j$, $N_j$ is the number of data points in class $j$ and $\bar{m}$ is the overall mean of all classes. The separation index $J$ is calculated for all bands (dimensions) of FM separately and shown in Fig. 1 across the center frequency the bands. The above analysis was repeated for FM features extracted using the smooth energy operator separation algorithm (SEOSA) [11], to verify that the results do not depend on the method of FM extraction. The results, seen in Fig. 1, show that the speaker separation is prominent around the range of frequencies from 700 Hz to 1100 Hz and 2200 Hz to 2800 Hz. This observation is consistent with both FM extraction methods. The higher $J$ measure observed around 3600 Hz is not considered significant for these data, since it is around the edge of the telephone bandwidth.

![Figure 1. Separation measure ($J$) for different bands in two different FM features, across the NIST2001 database](image)

3.2. KL distance

The Kullback-Leibler (KL) distance measures the distance between two probabilistic models in an information-theoretic sense. If the probabilistic model is a GMM then KL does not have a closed form solution and there are several approximations available to find the KL distance between two independently developed GMMs, such as methods based on Monte Carlo approximation and the unscented transform [12]. However, in our case the GMM model of each speaker is derived from maxim a posteriori (MAP) adaptation of only means from a common universal background model (UBM). An approximation for KL distance for GMMs adapted from a UBM is given in [12] as,

$$KL(f \parallel g) \approx \sum_{i=1}^{N} \alpha_i KL(f_i \parallel g_i)$$

(5)

where $f$ and $g$ are the two GMMs considered, $N$ is the total number of mixtures, $\alpha_i$ is the weight of the $i^{th}$ mixture.

![Figure 2. KL distance for different bands averaged over the NIST2001 database](image)

This approximation is valid as the GMM structure (mixture weights and variance) is consistent for mean-only MAP adaptation, which is the case in this paper. The symmetric version of the KL distance between two signal mixtures is [13]:

$$KL(f_i \parallel g_i) \approx 0.5 \left( \frac{\mu_{ij} - \mu_{ij}}{\Sigma_{ij}} \right)^T \left( \frac{1}{\Sigma f} + \frac{1}{\Sigma g} \right) \left( \frac{\mu_{ij} - \mu_{ij}}{\Sigma_{ij}} \right)$$
where $\Sigma$ is the covariance matrix, $\mu$ is the mean vector and $I$ is the identity matrix. The pairwise KL distances are calculated for each band (dimension) of the FM feature. These are averaged over all possible pairs in NIST 2001 training database. The variation of KL distance across the bands is shown in Figure 2. This analysis shows that the speaker separation is prominent around the range of frequencies from 300Hz to 700Hz and 2200Hz to 2800Hz. This observation is consistent across both FM extraction methods. The higher range (2200Hz to 2800Hz) is also consistent with the $J$ measure analysis.

4. Speaker-specific information at the classification level

Initially, a series of speaker recognition experiments were conducted by leaving out one band of the FM feature at a time (in each experiment only 13 dimensions out of 14 are used), followed by a single experiment using all bands of FM. By removing a band completely from the feature the previous assumption of dimensional independence is relaxed. The contribution to speaker recognition in terms of reduction in equal error rate (EER) is calculated by finding the difference between the EER obtained using all bands of FM and the EER obtained by leaving out one band of FM. These results, generated for NIST2001, indicate how speaker-specific information is distributed along different bands and are shown in Figure 3. Once again, these results are generally similar to observations based on the foregoing analyses using the separation index and the KL distance.

The observations derived from these feature, model and classification level analyses for the telephone bandwidth data can be summarized as: the frequency range roughly between (1) 300 Hz to 1000 Hz contributes to speaker recognition but not consistently in all three analyses (2) 1000 Hz to 2000 Hz contributes weakly to speaker recognition consistent with [4], (3) 2000 Hz to 3000 Hz contribute strongly to speaker recognition supporting the effect of the piriform fossa [2] and (4) 3000 Hz to 3700 Hz is undetermined as it crosses with the stop band region of the telephone bandwidth.

The latter observation for FM may be partly attributed to the cut off frequency of telephone bandwidth. Note that both observations are data-dependent.

Based on the results in Fig. 3, we hypothesized that if the bandwidths of the filters were reallocated to give more emphasis to the frequency ranges which contribute greater reductions in EER, then the performance of the entire feature should increase. Here, we consider 4 filter bank structures including 2 reallocation methods: (i) Uniform scale, (ii) Constant-Q, (iii) Empirical EER-weighted reallocation 1 and (iv) Empirical EER-weighted reallocation 2.

The method (iii) above is achieved by changing the EER reduction curve into a weighting function for the frequencies as below:

1) The EER reduction curve is offset so that the minimum point of the curve has a weight of 1.
2) The shifted curve, $E(f)$, is taken as the weighting function for re-allocation of filter bandwidth.
3) The area $(A)$ under $E(f)$ is equally divided to obtain the new bandwidths and the center frequency is the center within that bandwidth.
4) The lower and the upper cutoff frequencies of the $i^{th}$ filters, $f_{L}^{i}$ and $f_{U}^{i}$ are defined as in (7) and (8),

$$\sum_{f_{L}^{i}}^{f_{U}^{i}} E(f) = A$$

$$f_{L}^{i} = f_{L}^{i-1}$$

Method (iv) is achieved by retaining the center frequency as in (iii) and changing the bandwidth in (iii) to the uniform filter bandwidth (243Hz) if it is less. The bandwidth and center frequencies of all the filter banks are given in Figure 4.

![Figure 3. Reduction in EER in a series of 'leave one out experiments' using all-pole FM features, for the NIST2001.](image)

Observations from a comparison of the distribution of speaker-specific information obtained for MFCCs in [4] with Figure 3 for FM are: (1) both reveal that the contribution to speaker recognition is high in the area around 500Hz, drops beyond this to a minimum around 1100Hz for MFCCs [4] and around 1800 Hz for FM (Figure 3), then increasing to 2700Hz; (2) the contribution drops significantly after around 2700 Hz for FM, contrary to a continuous increase for MFCCs in [4].

![Figure 4. Bandwidth and center frequencies of the filter banks](image)

The performances of the FM based system with the above four filter banks in addition to the conventional Bark scale filters are given in Table 1 for the NIST 2001 database. The above FM system was combined with an MFCC-based system using score level fusion and the results are also included in Table 1. Then, in order to see the effect on a larger contemporary database, the male database of NIST 2006 is used to repeat for four filter bank structures and the results are given in Table 2. The baseline performances of MFCC-based system in terms of EER and detection cost function (DCF) are 7.6% and 3.5 respectively for NIST 2001 and 4.5% and 2.3 respectively for NIST 2006. Due to the extensive nature of the experiment a simple UBM-GMM (universal background model – Gaussian mixture model) system with feature warping as normalization is used for both MFCC and FM systems for the NIST 2001 database and for FM systems in the NIST 2006 male database while a near state of the art GMM-SVM (GMM-support vector machine) system with nuisance
attribute projection (NAP) and TNorm (test normalization) as normalization [14] was used for the MFCC system of NIST 2006 male database.

The significant improvement with the reallocated filter bank over the auditory motivated Bark scale filters in NIST 2001 database supports the observed patterns of the spread of speaker specific information seen in Figures 1 to 3. Though small, consistent improvements are observed when filters are reallocated except the DCF in NIST 2006 for the fused MFCC and FM system. Further, it shows that the auditory motivated filters are not the optimum choice for speaker recognition using FM features. Another interesting observation is the improvement across the size of the database. The improvement due to reallocation in terms of EER is 2.75% in NIST 2001 but 0.45% in NIST 2006 male database while the size of NIST 2001 is smaller than NIST 2006.

Table 1. Speaker recognition performance using NIST 2001 database in terms of EER (%), left and min DCF. (x100).

<table>
<thead>
<tr>
<th>Filter bank for FM</th>
<th>FM</th>
<th>FM + MFCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bark scale</td>
<td>12.7</td>
<td>5.36</td>
</tr>
<tr>
<td>Constant Q</td>
<td>13.2</td>
<td>5.99</td>
</tr>
<tr>
<td>Uniform scale</td>
<td>10.45</td>
<td>4.71</td>
</tr>
<tr>
<td>Reallocation 1</td>
<td>10.26</td>
<td>4.55</td>
</tr>
<tr>
<td>Reallocation 2</td>
<td>9.96</td>
<td>4.50</td>
</tr>
</tbody>
</table>

Table 2. Speaker recognition performance using NIST 2006 male database in terms of EER (%), left and min DCF. (x100).

<table>
<thead>
<tr>
<th>Filter bank for FM</th>
<th>FM</th>
<th>FM + MFCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bark scale</td>
<td>13.7</td>
<td>6.22</td>
</tr>
<tr>
<td>Uniform scale</td>
<td>13.7</td>
<td>6.08</td>
</tr>
<tr>
<td>Reallocation 1</td>
<td>13.5</td>
<td>5.98</td>
</tr>
<tr>
<td>Reallocation 2</td>
<td>13.25</td>
<td>5.97</td>
</tr>
</tbody>
</table>

Figure 5. Performance of speaker identification using all-pole FM features, for CHAINS database

Finally in order to understand the distribution of speaker-specific information in wideband speech, the CHAINS corpus [8] was employed (sampling frequency 44.1 kHz). In this analysis, the total bandwidth was divided into 80 non-overlapping uniform bands between 100Hz to 22000Hz, with a bandwidth of around 274 Hz. A series of experiments were conducted using features comprising a block of 5 (total bandwidth of 1370 Hz) and 10 (total bandwidth of 2737 Hz) consecutive bands at a time, moved at the rate of one band per experiment. The results are given in Fig. 5, and show that the bands from around 2000 Hz to 5000 Hz contribute more to the speaker identification task, which supports the effect of the hypopharynx [2]. A significant drop of speaker specific information is observed after about 12 kHz. So based on the above analyses, previous experimental analysis [4] and psychophysical explanations [2-4] it could be concluded that for effective speaker recognition the bandwidth of the database needs to be at least 8 kHz. Telephone bandwidth or 22.05 kHz bandwidth speech produces a loss in useful speaker recognition information or introduces more redundancy respectively.

5. Conclusions

This paper has studied the frequency distribution of speaker-specific information for frequency modulation features at the feature, model and classification level mainly for telephone data. For telephone bandwidth speech, the frequency range roughly between 2 to 3 kHz consistently provides the maximum speaker-specific contribution across all experiments, while for wideband speech the range extends up to 5 kHz. Further the auditory motivated Bark scale filter has been shown to be not the optimum filter for speaker recognition based on FM feature. Future work will examine further optimization of the filter banks based on the above observations.

6. References