DESIGNING A SPEAKER-DISCRIMINATIVE ADAPTIVE FILTER BANK FOR SPEAKER RECOGNITION

Tomi Kinnunen

Department of Computer Science
University of Joensuu, Finland
tkinnu@cs.joensuu.fi

ABSTRACT

A new filter bank approach for speaker recognition front-end is proposed. The conventional mel-scaled filter bank is replaced with a speaker-discriminative filter bank. Filter bank is selected from a library in adaptive basis, based on the broad phoneme class of the input frame. Each phoneme class is associated with its own filter bank. Each filter bank is designed in a way that emphasizes discriminative subbands that are characteristic for that phoneme. Experiments on TIMIT corpus show that the proposed method outperforms traditional MFCC features.

1. INTRODUCTION

Several studies have indicated that different phonemes have unequal discrimination powers between speakers [3, 10, 12]. That is, the inter-speaker variation of certain phonemes are different from other phonemes. For instance, in [3] vowels and nasals were found to be most discriminating phoneme groups.

Discrimination analysis of speech sounds can be, however, carried out from a non-phonetic viewpoint also. In several engineering-oriented studies, evidence of the different discrimination properties of certain frequency bands have been discovered [6, 14, 15]. For example, in [6] the spectra of speech were divided into upper and lower frequency regions with the cutoff frequency being the varied parameter. It was found, among other observations, that regions 0-4 kHz and 4-10 kHz are equally important for speaker recognition.

In [11] a more detailed analysis of frequency band discrimination was performed. Spectral analysis was carried out with a filter bank with triangular overlapping filters. Discrimination powers of these subbands were then evaluated with three different criteria, the *F-ratio* [1] being one criterion. A non-linear frequency warping based on the discrimination values was then proposed: more filters with narrower bandwidths were placed in the discriminative regions, while less filters with broader bandwidth were placed in the non-discriminative regions. The proposed system outperformed conventional mel-frequency warped filter bank.

Although the phonetic studies indicate differences in phoneme-level discrimination powers, no segmentation is usually done prior to discrimination analysis with the engineering-oriented approaches. The problem is, however, that when all different phoneme classes' data are pooled together, some discriminative frequency bands that are characteristic for a certain phoneme may be averaged away. The frequency of occurence of phonemes reflects directly to the discrimination values. As a consequence, if the corpus used in experiments contains a discriminating phoneme which is infrequent, its significance decreases.

In this work, we introduce an approach which falls in the middle ground between the ”phonetical”- and ”engineering”-oriented discrimination analyses.

Idea of the proposed front-end is illustrated in Fig. 1. Each speech frame is processed with a filter bank which is selected from a library of filter banks according to the phoneme class of the frame. Thus, each phoneme class is filtered in a customed way instead of a global filter bank as in [11].

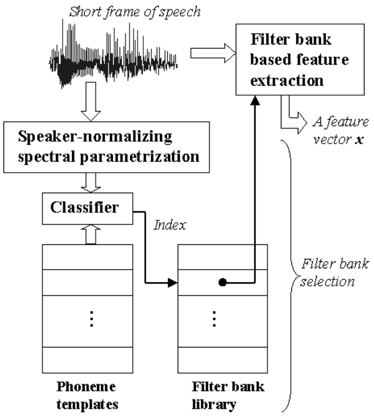


Figure 1: The idea of adaptive filter bank

The basic idea of the proposed method is simple. However, there arises immediately the following design issues:

- Which parametrization to choose in the determination of the phoneme class,
- How to generate and represent the phoneme templates,
- What is “optimal number” of the phoneme templates,
- How to compute discriminative values for subbands in phoneme-depended filter banks,
- How to exploit the filter bank in the feature extraction.

These are the substantial topics of this paper.

2. THE PHONEME CLASSIFIER

2.1 Representation of the phoneme templates

In order to be of general use, the phoneme template model must be speaker (or even language) independent. That is, the same model for all speakers can be used to find the phoneme classes. We denote this model as the *universal phoneme model* (UPM). Due to the requirement of speaker-independence, the UPM must be designed such that it accounts the differences between speakers and other sources of variability.

Note in Fig. 1 the block labeled "speaker-normalizing spectral parametrization". *Speaker normalization* means that we wish to de-emphasize speaker-depended features. We use the following parametrization which is general in speech recognition [12]:

- High emphasis with $H(z)$=1 - 0.97z^{-1},
- Frame length 30 ms, Hamming-windowed and shifted by 20 ms (33 % overlap),
- 12 lowest mel-frequency coefficients (MFCC), 20 triangular filters in the bank, coefficient c_0 discarded,
- Cepstral coefficients weighted by *raised sine* function.

2.2 Generation of the templates

We use clustering techniques [4, 5, 7] for generating the UPM from MFCC vectors. We use 100 speakers from the TIMIT corpus [9] as the training set. For each speaker, we take five speech files in the training data. These are downsampled to 8 kHz and processed with the parametrization given above. Final training set consists of approximately 100,000 vectors.

From the training set, a codebook is generated by the RLS algorithm [4]. The following different codebook sizes K are used: K=4, 8, 16, 32, 64.

The worth noticing point here is that we use *unsupervised learning* in the UPM generation; i.e. we do not use any explicit segmentation of speech or labeling of phonemes, since we are not interested in decoding the linguistic message of the input signal.

2.3 Classification of a frame

When applying the UPM in the phoneme classification, the class is simply determined by the nearest neighbor rule. Frame is first parametrized in the way described in Section 2.1, resulting in a single MFCC vector $\boldsymbol{x}$. The label of the phonetic class is then given by

$$i^* = \underset{\boldsymbol{p} \in UPM}{\arg\min}\ d(\boldsymbol{x}, \boldsymbol{p}), \qquad (1)$$

where d is the squared Euclidean distortion measure. The index i^* is sent to the filter bank library to select the associated filter bank (see Fig. 1).

3. DESIGNING THE LIBRARY OF DISCRIMINATIVE FILTER BANKS

3.1 Subband processing

We want to assign discriminative values for each subband per each phoneme class present in the UPM. To get started, we must specify what we mean here by a subband.

As in general speech processing front-ends [2] we use overlapping triangular filters to cover the entire frequency range (see Fig. 2). Filters are uniformly spaced and overlap by 50%. In this phase we wish to avoid using any nonlinear frequency warping, such as mel or Bark-scales, in order to be sure that each subband has equal contribution in the discrimination analysis.

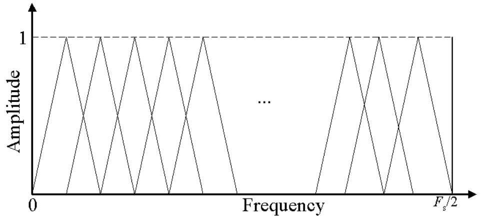


Figure 2: Uniform triangular filter bank

For a Hamming-windowed frame s, an N-point FFT $S[k]$ is first computed. The magnitude spectrum in dB-scale is then computed as $10\log_{10}|S[k]|$. The dB magnitude spectrum is weighted by the triangular filterbank of M filters, thus implying M subband energies E_{j}, j=1,...,M. These are collected in a M-dimensional vector $\boldsymbol{E}$ = $(E_1,...,E_{\mathrm{M}})^{\mathrm{T}}$. Hereafter, by "$j$th subband" we simply refer to E_{j}. We fix the number of filters to M=40. Thus, for the speech with sampling rate F_{s} = 8 kHz, the bandwidth of each filter is 100 Hz.

3.2 Assigning the discrimination values to subbands

We use the *F ratio* [1] for assigning a discrimination value for the jth subband of ith phoneme:

$$F_{i,j} = \frac{\text{Variance of speaker means of subband } j \text{ of phoneme } i}{\text{Average intraspeaker variance of subband } j \text{ of phoneme } i}. \qquad (2)$$

If the inter-speaker variability is large while inter-speaker variability being low, F ratio is large.

Since we wish to assign the F ratios for each phoneme-subband pair, we must first segment the training data into phonetic classes using the UPM described in Section 2. Then, for each "phonetic class pool" (i) we can compute the discrimination values for subbands (j) using F-ratio (2). The segmentation of the data into the pools is outlined in the following pseudocode.

```
Inputs:
•  Frames of speech {$s_{k,t}$}, where $k$ is the speaker number and $t$ is the frame number,
•  The UPM of $K$ phoneme templates, $UPM=\{p_1,...,p_K\}$,
•  The number of subbands, $M$
Outputs:
•  $K$ pools of subband vectors

Procedure:
FOR EACH speaker $k$ DO
    •  Get next frame $s = s_{k,t}$ ;
    •  Compute the MFCC vector $x$ for $s$ (as described in Section 2.1) ;
    •  Compute the subband vector $E$ for $s$ (as described in Section 3.1) ;
    •  Find nearest phoneme $i^*$ for $x$ from UPM by (1) ;
    •  Add the vector $E$ to the pool of  phoneme $i^*$ ;
ENDFOR
```

Figure 3: Segmentation of the training data for discrimination analysis

To put it in words, each frame is classified by its phonetic content, the UPM code vectors {$\boldsymbol{p}_i$} serving as the phoneme class representatives. The subband vector of the frame is assigned to the best matching phoneme template.

After the pooling, F ratios of each pool are computed. Indeed, different phonemes have different F curves as seen in Fig. 4, where we have used an UPM of size K=8.

A few prelimary observations can be made from the F curves. Firstly, nearly all phonemes have a peak in discrimination values approximately in the subbands 2-4, which correspond to frequency range 50-250 Hz. Secondly, one may see some resemblance of the F ratio shapes with the envelopes of smoothed LPC spectra, thus indicating the importance of formant structure and overall spectral envelope in speaker recognition.

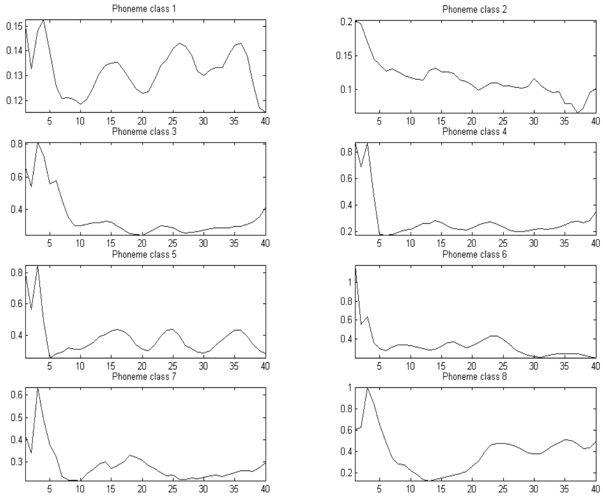


Figure 3: F ratios of subbands for different phonemes (UPM size K=8)

We run also an experiment in which, instead of pre-smoothing the spectra with a filter bank, we used all the magnitude values from FFT analysis as such and computed the F ratios. We found soon out that the F curves obtained in this way were very noisy; further, the computational load for this method is huge compared pre-smoothing using the filter bank. For these reasons, we end up using the filter bank.

3.3 Utilization of the filter bank in feature extraction

Once the F ratios are computed for each phoneme-subband pair, it is straightforward to utilize them in the feature extraction. The broad phoneme class i^* is first found by (1). This is followed by the subband analysis as described in Section 3.1, leading to vector $\boldsymbol{E}$. The components of $\boldsymbol{E}$ are then weighted by the relative F ratio of the subband:

$$E'_j = E_j \frac{F_{i^*,j}}{\sum_{m=1}^{M} F_{i^*,m}} \tag{3}$$

An example of subband weighting is shown in Fig. 4. The figures from top to down show the magnitude spectrum, filtered magnitude spectrum, relative weights for each subband, and the weighted filter outputs.

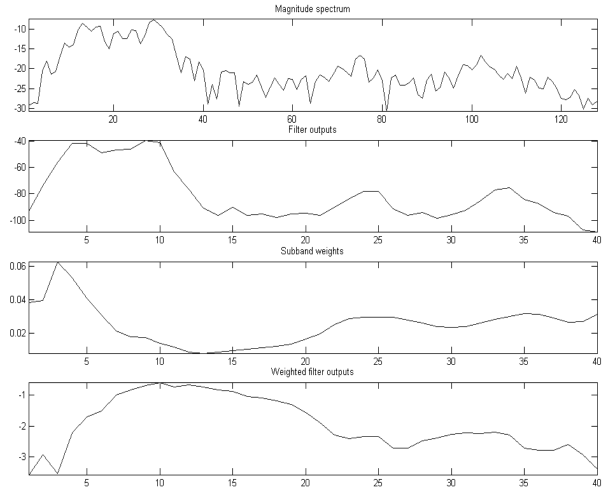


Figure 4 : An Example of subband weighting

Weighted filter outputs are then fed to discrete cosine transform (DCT) for decorrelating the features. Only the lowest L coefficients of DCT are retained, excluding the 0th coefficient.

In summary, the processing steps are same to that of the conventional MFCC analysis, except for that the mel-spaced filter bank is replaced with the discriminative filter bank. Hereafter, we abbreviate the features obtained in this way by *ADFB-cep* (standing for *Adaptive Discriminative Filter Bank Cepstrum*).

4. RESULTS

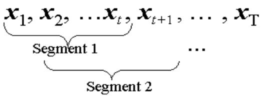
The overall process of evaluating the proposed approach consists of the following steps:

- Create UPM as described in Section 2 (Using speaker set *SET 1*),
- Use independent data for finding the F ratios as described in Sections 3.1 and 3.2 (*SET 2*),
- Using third speaker set (*SET 3*), compute the *ADFB-cep* features as described in Section 3.3. We choose the number of filters to M=40 and number of coefficients to L=20. *SET 3* is further divided into training and evaluation sets.

All the three sets are disjoint. In this way we ensure that results will be not biased by the tuning to the training set; that is, we wish to have a general front-end without the need to construct the UPM and/or the filter design data each time the database is switched.

Each of the three sets consist of 100 speakers. We use VQ codebooks as speaker models [8, 13], each model having 64 code vectors and created using the RLS clustering method [4]. Average duration of the training speech data is 15 seconds.

Each test set $X = \{\boldsymbol{x}_1,...,\boldsymbol{x}_T\}$ is divided into overlapping segments as shown in the following:


$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_t, \boldsymbol{x}_{t+1}, \ldots, \boldsymbol{x}_T$
Segment 1
…
Segment 2

Average duration of the test segment is about 1 second. Each of the segments is classified using the speaker models by the minimum average quantization error rule [13]. We use the percentage of correctly classified segments as the evaluation criterion.. The results for different UPM sizes are shown in Table 1.

Table 1: Evaluation results

UPM size	ID rate (%)
4	69.37
8	74.85
16	67.071
32	58.49
64	55.73

For comparison, conventional 20 mel-cepstral coefficients (MFCC) were computed with same frame rate and equal parameters: number of mel-filters was 40 and the number of coefficients was 20. The identification rate using MFCC was 61.96.

Based on these experiments, we make several observations. Firstly, the optimum size of UPM is K=8. When the UPM size is increased, results get poor. Also, the differences in performance are quite large, which suggests that we should use a linear scale instead of exponential when finding the "optimum size". Secondly, and more interestingly, the proposed method outperforms MFCC parameters, even if the UPM size is not "optimal". The overall identification rates are quite poor in all cases, due to the very short test segment length.

5. CONCLUSIONS

A new feature set based on discriminative weighting of the characteristic subbands for each "phoneme class" was proposed and evaluated experimentally. Prelimary results are very encouraging since they outperform the popular MFCC features. In future experiments, we plan to include cross-language evaluation, careful optimization of the UPM, and other discrimination criteria in addition to F ratio.

6. REFERENCES

[1] Campbell, J., "Speaker Recognition: A Tutorial," *Proc. IEEE*, **85**(9): 1437-1462, 1997.

[2] Deller, J.R. Jr., Hansen, J.H.L. and Proakis, J.G., *Discrete-time Processing of Speech Signals*. Macmillan Publishing Company, New York, 2000.

[3] Eatock, J.P. and Mason, J.S., "A Quantitative Assessment of the Relative Speaker Discriminating Properties of Phonemes", *Proc. ICASSP'94*: 133-136, Adelaide, 1994.

[4] Fränti, P. and Kivijärvi, J., "Randomized Local Search Algorithm for the Clustering Problem", *Pattern Analysis and Applications*, **3**(4): 358-369, 2000.

[5] Gersho, A. and Gray, R.M., *Vector Quantization and Signal Compression*, Kluwer Acad. Pub., 1992.

[6] Hayakawa, S. and Itakura, F.: "Text-Dependent Speaker Recognition Using the Information in the Higher Frequency Band", *Proc. ICASSP'94*: 137-140, Adelaide, 1994.

[7] Jain A.K, Murty M.N. and Flynn P.J., "Data Clustering: A Review", *ACM Computing Surveys* **31**(3): 264-323, 1999.

[8] Kinnunen, T., Kärkkäinen, I., "Class-Discriminative Weighted Distortion Measure for VQ-Based Speaker Identification", accepted for publication in *Proc. Joint IAPR International Workshop on Statistical Pattern Recognition (SPR 2002)*, Windsor, August 6-9, 2002.

[9] *Linguistic Data Consortium*, http://www.ldc.upenn.edu/

[10] Nolan, F., *The Phonetic Bases of Speaker Recognition*, Cambridge CUP, 1983.

[11] Orman, Ö.D. and Arslan, L.M., "Frequency Analysis of Speaker Identification", *2001 Speaker Odyssey*, Jerusalem, 2001.

[12] Rabiner, L. and Juang, B.-H., *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[13] Soong, F.K., Rosenberg, A.E., Juang, B.-H. and Rabiner, L.R.: "A Vector Quantization Approach to Speaker Recognition", *AT&T Technical Journal*, **66**: 14-26, 1987.

[14] v. Vuuren, S. and Hermansky, H., "On the Importance of Components of the Modulation Spectrum for Speaker Verification", *Proc. ICSLP '98*: 3205-3208, Sydney, 1998.

[15] Yoshida, K., Takagi, K. and Ozeki, K., "Speaker Identification Using Subband HMMs", *Proc. EUROSPEECH '99*: 1019 - 1022, Budapest, 1999.