Proceedings of the Institute of Acoustics

SPEECH RECOGNITION USING SPEAKER DEPENDENT FREQUENCY WARPING

S. Flesch and M. Brookes

Signal Processing Section, Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, London SW7 3BT, UK (http://www.dsp.ee.ic.ac.uk)

1. ABSTRACT

The spectral differences between examples of a particular phoneme uttered by different individuals may in part be modelled as a smooth non-linear warping of the frequency axis. This paper describes a speech recognition system in which the peak frequencies of the analysis filter bank are optimised for each speaker instead of being fixed at predetermined mel-spaced increments as is usually the case. In the proposed system, the outputs from the speaker-specific filter bank are converted to cepstrum coefficients to form the parameter vector for a conventional speaker-independent hidden Markov model recogniser. The appropriate filter bank frequencies to use for a new speaker are determined by means of a gradient descent algorithm that minimises the errors in a phoneme classification task. The use of such speaker-dependent frequency warping is applied to a phoneme classification experiment using a subset of the TIMIT database. The application of this technique to speaker-adaptive recognition systems is also discussed.

2. INTRODUCTION

Differences between speakers generally cause speaker-independent recognition systems to have lower performance than speaker-dependent systems. Many researchers have proposed ways of reducing this performance gap by compensating for these inter-speaker differences during recognition. The proposed compensation techniques generally follow one of two approaches. In the first approach, an adaptive transformation is applied to a set of speaker-independent models to improve their fit to the input speech. In the second approach, a normalising transformation is applied to the input speech during both training and recognition. The procedures described in this paper follow the second of these approaches.

One of the significant sources of inter-speaker variability arises from differences in vocal tract length. Acoustic theory indicates that, in the absence of any compensation by the speaker, a linear scaling of the vocal tract dimensions will result in a linear scaling of formant frequencies. Accordingly, a number of papers have suggested warping the frequency scale of the input speech in a linear [1], piecewise linear [2] or non-linear [3] manner. In each case the form of the warping function is fixed in advance and a single parameter is varied to compensate for inter-speaker differences.

The normalising procedure described in this paper differs from the previous approaches in two ways: it makes no prior assumptions about the shape of the warping function and it is determined by minimising the error in a phoneme classification task rather than using the more common maximum likelihood criterion.

3. SPEECH RECOGNISER STRUCTURE

The speech recogniser used in these experiments has a conventional structure in which input speech frames are converted to cepstral parameters which form the input of a hidden Markov model recogniser.

In the front end of the recogniser, input speech is divided into overlapping 16ms frames which are windowed and converted into the frequency domain by a discrete Fourier transform. The resultant power spectrum is then filtered by a set of bandpass filters and converted to the cepstral domain by a discrete cosine transform (DCT). This process may be represented in matrix form as

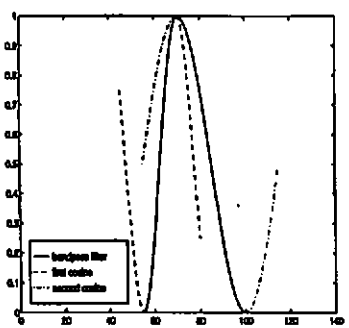
$$\mathbf{o} = \mathbf{T}\, log(\mathbf{H}\, sqr(\mathbf{FWx})) \tag{1}$$

in which the column vectors $\mathbf{x}$ and $\mathbf{o}$ represent the input speech frame and the cepstral coefficients, and the matrices W, F, H and T represent the windowing, FFT, filter bank and DCT. The functions log() and sqr() denote element by element log and square functions.

In most recognisers, the matrix H defines a fixed set of mel-spaced bandpass filters whose response in the power spectral domain has an asymmetric triangular shape. The distinctive feature of the work presented in this paper is that the centre frequencies of the filters defined by H are chosen for each speaker to optimise recognition performance.

For reasons of computational and analytical convenience, we have used raised cosine rather than triangular filters. As with the conventional mel filter bank [4], filters must be asymmetric to ensure that each column of H sums to unity. If the filter bank centre frequencies are denoted by the vector b, the matrix H may be defined by:

$$\mathbf{H}[k,l] = \begin{cases} \frac{1}{2} + \frac{1}{2}\, cos\Big(\pi\,(l - \mathbf{b}[k])/(\mathbf{b}[k] - \mathbf{b}[k-1])\Big) \\ \quad for\ \ \mathbf{b}[k-1] \le l < \mathbf{b}[k] \\ \frac{1}{2} + \frac{1}{2}\, cos\Big(\pi\,(l - \mathbf{b}[k])/(\mathbf{b}[k+1] - \mathbf{b}[k])\Big) \\ \quad for\ \ \mathbf{b}[k] \le l \le \mathbf{b}[k+1] \\ 0 \quad for\ \ l < \mathbf{b}[k-1]\ or\ l > \mathbf{b}[k+1] \end{cases} \tag{2}$$



4. MINIMUM CLASSIFICATION ERROR

The optimisation of the error count of a phoneme classification task was first introduced by [5, 6] as a new discriminative training procedure. Recently, the minimum classification error method (MCE) was introduced as a new hidden Markov model (HMM) training algorithm based on the generalised probabilistic descent (GPD) method [7, 8, 9]. In this method, the classification error cost function is defined as a function of the HMM parameters. Then, the gradient of the cost function is calculated and a probabilistic descent method is applied in order to find the optimal HMM parameters. In our work, we have expressed the classification error cost function as a function of the filter bank peak frequencies but have used a conventional gradient descent procedure to find optimal peak frequencies for a particular speaker.

Let $\mathbf{O}^l = \{\mathbf{o}^l_1, \mathbf{o}^l_2, ..., \mathbf{o}^l_{T_l}\}$ denote the l-th training sequence in the cepstral domain, where T_l is the number of frames. Each training sequence belongs to a class $k = \{1, 2, ..., K\}$, modelled by a HMM Λ_k, $\Lambda_k \in \Lambda$ where $\Lambda = \{\Lambda_1, \Lambda_2, ..., \Lambda_K\}$ represents the whole set of HMMs. Let the observation density of an HMM be Gaussian with mean μ and diagonal covariance C, and the transition probabilities be stored in matrix a.

The formulation of the cost function consists of four functions:

4.1 *Discriminant function.*
This function is a measure of the distance between an input training sequence and the corresponding HMM from list Λ. The negative log-likelihood score is chosen in our study.

$$g_k(\mathbf{O}^l, \Lambda) = \frac{1}{T_l}\left(\frac{1}{2}\sum_{t=1}^{T_l}(\mu^k_{S(t)} - \mathbf{o}^l_{(t)})^T \mathbf{C}^{k-1}_{S(t)}(\mu^k_{S(t)} - \mathbf{o}^l_{(t)}) - \sum_{t=1}^{T_l-1} log\, a_{S(t),S(t+1)} + \frac{1}{2}\sum_{t=1}^{T_l} log(det(\mathbf{C}^k_{S(t)}))\right) \quad (3)$$

where, $S(t)$ represents the alignment sequence issued from the Viterbi alignment, and μ^k and $\mathbf{C}^k$ are the HMM output probability parameters of class k. The constant term has been omitted.

4.2 *Misclassification measure.*
The misclassification function measures the degree of confusion between the correct class and all other competing classes.

$$d(\mathbf{O}^l, \Lambda) = g_k(\mathbf{O}^l, \Lambda) + \frac{1}{\eta} log\left(\frac{1}{K-1}\sum_{j \neq k} exp(-\eta g_j(\mathbf{O}^l, \Lambda))\right) \quad (4)$$

where k is the class to which $\mathbf{O}^l$ belongs. When $\eta \to \infty$, this becomes:

$$d(\mathbf{O}^l, \Lambda) = g_k(\mathbf{O}^l, \Lambda) - \min_{j \neq k} g_j(\mathbf{O}^l, \Lambda) \quad (5)$$

which is the distance of the input observation between the correct model and the best incorrect model (this simplification is applied in our research). Clearly, a negative value of $d(\mathbf{O}^l, \Lambda)$ indicates a correct classification and a positive value indicates a misclassification.

4.3 *Smooth loss function.*
This function weights the misclassification measure using a sigmoid function. When an input is correctly classified ($d(\mathbf{O}^l,\Lambda) << 0$), the weight (or penalty) will be near 0. Likewise, a misclassified input will have a near unity weight.

$$\gamma(\mathbf{O}^l,\Lambda) = \frac{1}{2}(1+tanh(\beta\, d(\mathbf{O}^l,\Lambda))) \qquad \beta > 0 \tag{6}$$

As $\beta \rightarrow \infty$, the sigmoid function tends to a step function.

4.4 *Cost function.*
The cost function is as follows:

$$D(\Lambda) = \sum_{l=1}^{L} \gamma(\mathbf{O}^l,\Lambda) \tag{7}$$

$D(\Lambda)$ represents a measure of the misclassification of the given observation sequences ($\mathbf{O}^1 \ldots \mathbf{O}^L$) for the entire classifier Λ. When the discriminant function is the negative log-likelihood score and the loss function is a step function, this cost function is precisely the error rate.

5. FILTER BANK AND HMM OPTIMISATION

The training phase of our speaker normalised automatic speech recogniser (ASR) jointly optimises a set of speaker-specific filter banks and the phoneme models within the classifier. The test phase derives the optimal filter bank of a new speaker for the fixed classifier estimated during the training phase. In both phases, the filter bank peak frequencies are derived by minimising the previously described cost function along its gradient.

5.1 *Gradient Analysis.*
The cost function evaluated for a specific speaker is a non linear function of the vector of peak frequencies. The gradient ∇D is defined by differentiating the cost function $D(\Lambda)$ with respect to each peak frequency:

$$\nabla D = [\frac{\delta D}{\delta \mathbf{b}[0]}, \frac{\delta D}{\delta \mathbf{b}[1]}, \cdots, \frac{\delta D}{\delta \mathbf{b}[N_{filt}+1]}]^T \tag{8}$$

where each element of the gradient vector is calculated by applying the chain rule:

$$\frac{\delta D}{\delta \mathbf{b}[i]} = \frac{\delta D}{\delta \gamma} \cdot \frac{\delta \gamma}{\delta d} \cdot \frac{\delta d}{\delta \mathbf{O}^l} \cdot \frac{\delta \mathbf{O}^l}{\delta \mathbf{H}} \cdot \frac{\delta \mathbf{H}}{\delta \mathbf{b}[i]} \tag{9}$$

The above expression includes the derivative of the loss function,

$$\frac{\delta \gamma(\mathbf{O}^l,\Lambda)}{\delta d} = \frac{1}{1+cosh(2\beta d(\mathbf{O}^l,\Lambda))} \tag{10}$$

This bell shaped derivative function directs the gradient search by concentrating on the improvement of training tokens near the decision threshold and leaving out well recognised or hopelessly recognised tokens.

The differentiation of the cost with respect to one peak frequency, expressed by (9), is easily obtained. However, the detailed calculation is too long to be presented in this paper.

5.2 *Test Phase.*

The filter bank peak frequencies of a new speaker are determined by applying a steepest descent algorithm to the cost function. Each iteration of this algorithm updates the peak frequencies by a small amount along the gradient direction. Since the filter bank varies after each iteration, it is necessary to recalculate the Viterbi alignment of each test sequence and reevaluate the cost function. The following iterative procedure is applied:

1. Determine the most probable incorrect model for each training token
2. Evaluate the objective function and its derivative with respect to each element of the speaker's filter bank.
3. Adjust the filter bank in the direction that will decrease the objective function.
4. Loop to step 2 a number of times (e.g. 2)
5. Loop to step 1 until convergence occurs

5.3 *Training Phase.*

In the training phase, the filter banks of all training speakers are optimised and speaker-independent phoneme models are estimated. The search for the optimal front end is performed by applying the test phase iterative procedure for all training speakers at the same time. In order to find the optimal classifier an additional loop is added to the previous process, resulting in the following iterative procedure:

1. Initialise all filter banks
2. Use standard HMM training to create phoneme models
3. For each speaker:
   4. Determine the most probable incorrect model for each training token
   5. Evaluate the objective function and its derivative with respect to each element of the speaker's filter bank.
   6. Adjust the filter bank in the direction that will decrease the objective function.
   7. Loop to step 5 a number of times (e.g. 2)
   8. Loop to step 4 a number of times (e.g. 4).
9. Loop to step 2 until convergence occurs.

6. EXPERIMENT

The technique presented in this paper has been tested on a phoneme classification experiment of 49 speakers of the TIMIT database. The first dialect region *dr1* contains the speech data of 38 speakers for training purpose and 11 speakers for testing the system. In the training phase, eight sentences are used per speaker to derive the optimal front end and the classifier models. In the test phase (speaker-dependent recognition), eight sentences are used per speaker to derive the optimal front end for a fixed classifier. In each phase, the recognition rate is derived from the same 8 sentences used during the optimisation stage.

The speech waveform is segmented into 16 ms frames every 8 ms and passed through a preemphasis filter. A 256 point discrete Fourier transform is subsequently performed. The resulting spectrum frame is normalised before being filtered through the speaker-specific filter bank which contains 12 raised cosine bandpass filters. A 12 to 10 point discrete cosine transform is applied to the log energy output of the filters in order to give a 10 cepstrum coefficient vector.

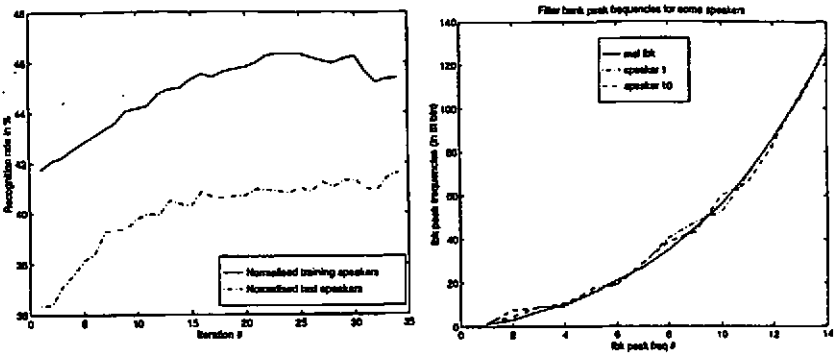
The pattern matching stage is implemented by 48, three-state, context-independent, left-to-right hidden Markov models representing the phoneme vocabulary. The state output probability distribution consists of a single Gaussian mixture with diagonal variance. The models are trained using the maximum likelihood objective function by applying iterations of the Baum-Welch algorithm on isolated models.

The recognition rate of a phoneme classification task is estimated at each iteration of the training procedure. The test procedure is performed on the speaker-independent models generated after each iteration of the training procedure. The resulting filter banks of the test speakers are used to calculate the recognition rate of a phoneme classification task performed on the test data.

7. RESULTS AND DISCUSSION

The results of the phoneme classification task after each iteration of the adaptation algorithm are shown in figure 1(a). The top curve represents the evolution of the recognition rate of the training set of speakers whereas the bottom curve represents the test set. The recognition rate of both training and test sets improves steadily. After 34 iterations, the recognition rate increased from 36.34% to 41.58%, a reduction in the error rate of 8%. Generally, the iterative adaptation improves recognition and seems to converge. However, sometimes the update on the HMM parameters results in a sharp rise in the cost function for specific speakers. This results in a drop in the general recognition rate as seen in figure 1(a). The speaker-specific cost function requires a few iterations to return to its normal decreasing course.

The starting point of our search for the optimal speaker-specific filter bank is the mel-spaced filter bank. The peak frequencies of two filter banks after optimisation are shown in figure 1(b). It is clear that these optimal filter banks are near the mel scale and a test can be carried out to establish if a different starting point will give different optimised filter banks.



(a) Recognition rates for training and test data

(b) New filter bank for two speakers

Figure 1: Filter bank optimisation of training and test speakers

8. CONCLUSIONS

We have proposed a technique based on a speaker-specific transformation that improved the recognition rate of a phoneme classification task. The novel aspect of our approach is to use the filter bank as a speaker-specific transformation. The MCE optimising criterion was used as a cost function of a gradient descent method in order to find the optimal peak frequencies. The front end and classifier of our ASR were optimised iteratively by applying the gradient descent. This adaptation algorithm was tested on a subset of the TIMIT database resulting in a 8% reduction of the error rate.

The proposed technique improves the recognition rate of a speaker-dependent transformation / speaker-independent classifier ASR. The filter banks of the test speakers were optimised by a gradient descent method as explained in section 5. The next step of the research currently being undertaken is to implement an unsupervised speaker adaptation. In such a system, the training phase builds a space of filter banks and a space of speaker characteristics derived from the speech data of training speakers. A mapping between both spaces is created for its use in the test phase. The test phase of the new system involves extracting the new speaker's characteristics and using a predetermined mapping that selects the optimal filter bank in the space of filter banks.

In order to tackle the local minimum burden of the the steepest descent method, an alternative optimisation technique is currently being examined in which the peak frequencies are optimised by means of a genetic algorithm. In our environment, the genetic algorithm has three advantages: the global minimum is found, the front end and the classifier are optimised simultaneously rather than sequentially and it can be applied

to any recognition task including phoneme classification, whole-word or sub-word continuous speech recognition. Early experiments have shown encouraging results.

9. ACKNOWLEDGEMENTS

We would like to thank Keith Ponting and Roger Moore, DRA Malvern, for their helpful comments and suggestions. This work is supported by an EPSRC and DRA grant.

10. REFERENCES

[1] L. LEE AND R. ROSE. "Speaker normalization using efficient frequency warping procedures". International Conference on Acoustics, Speech and Signal Processing, 1:353–356, 1996.

[2] S. WEGMAN, D. MCALLASTER, J. ORLOFF, AND B. PESKIN. "Speaker normalisation on conversational telephone speech". International Conference on Acoustics, Speech and Signal Processing, 1:339–341, 1996.

[3] E. EIDE AND H. GISH. "A parametric approach to vocal tract length normalization". International Conference on Acoustics, Speech and Signal Processing, 1:346–348, 1996.

[4] S. DAVIS AND P. MERMELSTEIN. "Comparison of parametric respresentations for monosyllabic word recognition in continuously spoken sentences". IEEE Transactions on Acoustics, Speech and Signal Processing, 28(4):357–366, August 1980.

[5] L. R. BAHL, P. F. BROWN, P. V. DE SOUZA, AND R. L. MERCER. "A new algorithm for the estimation of hidden Markov model parameters". International Conference on Acoustics, Speech and Signal Processing, S11.2:493–496, 1988.

[6] C. AYER. "Optimal linear transforms for speech recognition". PhD thesis, Electrical Engineering, Imperial College of Science , Technology and Medecine, 1992.

[7] B. JUANG AND S. KATAGIRI. "Discriminative learning for minimum error classification". IEEE Transaction on Signal Processing, 40(12):3043–3054, December 1992.

[8] W. CHOU, B. JUANG, AND C. LEE. "Segmental GPD training of HMM based speech recogniser". International Conference on Acoustics, Speech and Signal Processing, I:473–476, 1992.

[9] C. RATHINAVELU AND L. DENG. "Use of generalized dynamic feature parameters for speech recognition: maximum likelihood and minimum classification error approaches". International Conference on Acoustics, Speech and Signal Processing, I:373–376, 1995.