A FAST METHOD OF SPEAKER NORMALISATION USING FORMANT ESTIMATION

M. Lincoln[1] *S. Cox*[1] *S. Ringland*[2]

[1]University Of East Anglia, Norwich, Norfolk, UK
ml@sys.uea.ac.uk
[2]BT Laboratories, Martlesham Heath, Suffolk, UK
spar@saltfarm.bt.co.uk

ABSTRACT

It has recently been shown that normalisation of vocal tract length can significantly increase recognition accuracy in speaker independent automatic speech recognition systems. An inherent difficulty with this technique is in automatically estimating the normalisation parameter from a new speaker's speech and previous techniques have typically relied on an exhaustive search to estimate this parameter. In this paper, we present a method of normalising utterances by a linear warping of mel filter bank channels in which the normalisation parameter is estimated by fitting formant estimates to a probabilistic model. This method is fast, computationally inexpensive and requires only a limited amount of data for estimation. It generates normalisations which are close to those which would be found by an exhaustive search. The normalisation is applied to a phoneme recognition task using the TIMIT database and results show a useful improvement over an un-normalised speaker independent system.

1. INTRODUCTION

Speaker independent (SI) automatic speech recognition systems generally have a lower recognition accuracy than their speaker dependent counterparts. This is largely caused by variations both physiological (eg vocal tract size and shape) and psychological (e.g. mood, accent, speaking rate) between speakers, resulting in SI models having significantly larger variances. Previous work ([2],[4], [1]) has shown that the application of a frequency normalisation to the utterances of each speaker may reduce inter-speaker variability caused by differing vocal tract lengths, hence reducing model variances and increasing recognition performance if used in an SI-ASR system.

Previously, speaker normalisation factors have usually been found by an exhaustive search of the normalisation space (e.g [2],[4]), which is computationally unattractive. The method reported here uses a scaling of the frequency axis to fit estimates of the first and second formant frequencies of (labeled) sounds from the speaker to class conditional distributions of the formants. This scaling is then used to normalise all the data from the speaker.

2. DATA NORMALISATION PRIOR TO MODEL TRAINING

2.1. Formant Estimation From LP Parameters

Since computational efficiency is important to the normalisation procedure it was decided to use a computationally efficient method of formant picking. Linear prediction represents the vocal tract response as an all pole filter :

$$H(z) = \frac{1}{\sum_{i=1}^{m} a_i z^{-i}} \quad (1)$$

$$= \frac{1}{A(z)}, \quad (2)$$

where a_i are the predictor coefficients and m is the analysis order. The roots of the predictor polynomial, $A(z)$, can be found using a root finding algorithm [3] , and the roots can be used to provide candidate frequencies and bandwidths for the speech formants. For each root, r_i, the frequencies and bandwidths are given by:

$$F_i = \frac{\theta_i f_s}{2\pi} \quad (3)$$

$$B_i = \frac{-\ln |r_i| f_s}{\pi} \quad (4)$$

where f_s is the sampling frequency and θ_i and $|r_i|$ are the angle and magnitude of r_i respectively.

The roots representing the formants typically have very small bandwidths (by observation, typically $|r_i| > 0.9$, corresponding to $B_i = 537$Hz for $f_s = 16$KHz) and low frequencies. The first and second formants were estimated by sorting the roots into order of ascending frequency, and extracting the two lowest with $|r_i| > 0.9$.

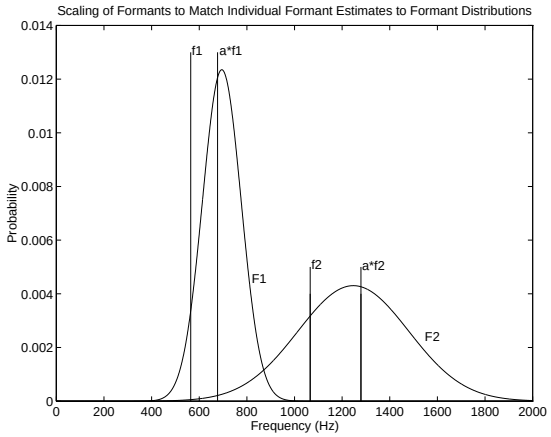

Figure 1. Normalisation Factor Estimation

2.2. Estimation of the normalisation factor for a speaker

The training-data was the complete training-set of the TIMIT database (426 speakers). Normalisation of the training-data is a two-stage process:

- Stage 1: F1 and F2 estimates for each frame of each vowel segment in the training-data are made and are used to estimate a uni-dimensional Gaussian distribution for each formant for each vowel-class.
- Stage 2: The estimates of F1 and F2 for the i'th frame given by a speaker (f_i^1 and f_i^2) are used together with the two distributions appropriate to the vowel-class of the frame to estimate a normalisation factor a_i for the frame. a_i is calculated as follows:

$$a_i = \arg\max_a \Pr(af_i^1 \mid F^1)\Pr(af_i^2 \mid F^2) \quad (5)$$

$$a_i = \arg\max_a \left(\frac{1}{2\pi\sigma_1\sigma_2}\exp -\frac{1}{2}\frac{(af_i^1-\mu_1)^2}{\sigma_1^2} \times \exp -\frac{1}{2}\frac{(af_i^2-\mu_2)^2}{\sigma_2^2}\right) \quad (6)$$

where F^1 and F^2 are the distributions for the vowel-class appropriate for the frame, with means μ_1, μ_2 and standard deviations σ_1, σ_2 (figure 1).

The closed form solution for a which maximises equation 6 is given by:

$$a_i = \frac{f_i^1\mu_1/(\sigma_1)^2 + f_i^2\mu_2/(\sigma_2)^2}{(f_i^1/\sigma_1)^2 + (f_i^2/\sigma_2)^2} \quad (7)$$

These normalisation factors are then combined to form a single overall normalisation factor estimate $a(I)$ for speaker I which is a likelihood-weighted combination of the individual a_i^I's:

$$a(I) = \frac{\sum_{i=1}^{N_v} a_i^I Pr(a_i f_i^1 \mid F^1)\Pr(a_i f_i^2 \mid F^2)}{\sum_{k=1}^{N_v} Pr(a_k f_k^1 \mid F^1)\Pr(a_k f_k^2 \mid F^2)} \quad (8)$$

where a_i is the normalisation for frame i, and N_v is the total number of vowel frames for speaker I.

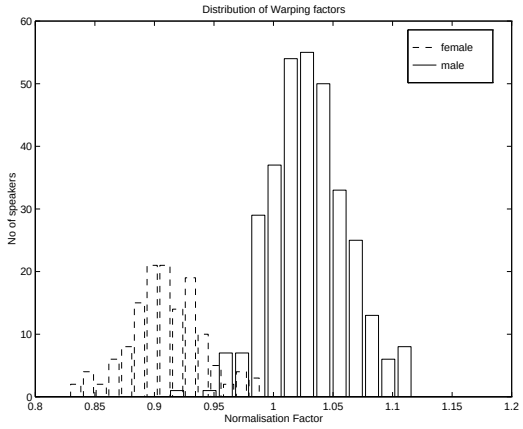

Figure 2. Distribution of warping factors

The distribution of normalisation factors is shown in figure 2. There is a clear distinction between the normalisations for male and female speakers. Female speakers generally have a normalisation factor less than one, while males have a factor greater than one. The normalisation compresses the frequency response of the female speakers and expands it for the males. This is what would intuitively be expected since, in general, women have shorter vocal tracts and correspondingly higher formants than men.

2.3. Filter bank normalisation

Once a normalisation factor $a(I)$ has been estimated for speaker I (as described in 2.2.), MFCC coefficients are generated for the speaker's data as follows:

1. A normalised filter bank is generated by scaling the centre-frequencies of a mel-scale filter bank according to the value of $a(I)$ (figure 3).
2. The LPC coefficients for each frame (already calculated for estimation of formant frequencies) are used to estimate a vocal tract frequency response by evaluating the filter response at 800 evenly spaced intervals across the sampling frequency.
3. This response is filtered by the normalised filter bank, the log of the filter bank output taken and a DCT applied to form the MFCCs

To ensure that accent variations between speakers did not mask the vocal tract effects, HMM phoneme

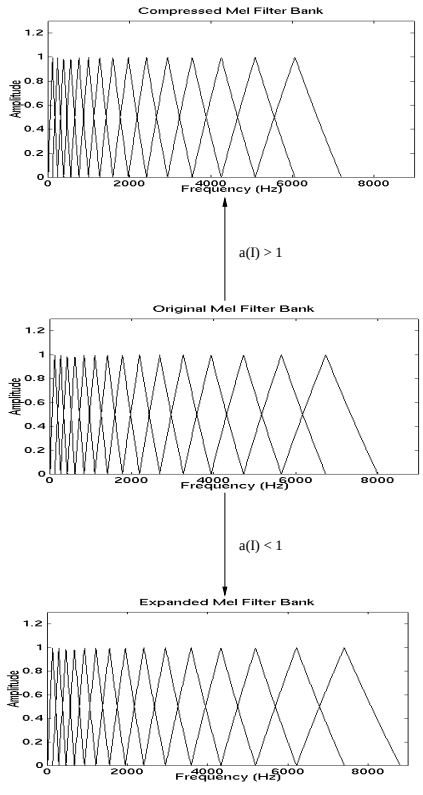

Figure 3. Filter Bank Warping

models for each of the 8 dialect regions represented in the TIMIT database were then built using the normalised data. A 3 state, left-right (no skips), single Gaussian, diagonal covariance matrix topology was used.

3. RECOGNITION EXPERIMENTS

Several recognition experiments were conducted in order to evaluate the effectivness of the normalisation. Initially, since only vowel segments of speech were used to calculate the normalisation factor, the normalised filter bank was only used to parameterise the vowel sounds. Other speech frames used a standard mel-scaled filter (figure 4) The test data was the complete test-set of the TIMIT database (197 talkers, 10 sentences per speaker) and was normalised using the same, 'vowel-only' normalisation procedure. Performance was 40.66% phoneme accuracy for the normalised case (averaged across the 8 dialect regions) compared with 38.95% using un-normalised data and models. To investigate the effectivness of the procedure on non-vowel sounds, normalisation was then performed on all the speech frames regardless of whether they had been used to evaluate the normalisation value. Recognition accuracy increased to 41.72% using fully normalised training and test data. Finally, in order to investigate the robustness of the normalisation factor, two sentences from each test talker were used to estimate the normalisation factor $a(I)$ and all ten sentences were then normalised using $a(I)$ and recognised using the appropriate dialect models. Using normalisation, performance was 41.74% phoneme accuracy.

These results for individual dialect regions in TIMIT are shown in figure 5.

Table 1. Results of recognition experiments

Normalisation Method	Recognition Accuracy
None	38.95%
Vowel-Only	40.66%
All	41.72%
Estimate Derived From Two Sentences	41.74%

4. DISCUSSION AND FUTURE WORK

In this preliminary investigation, we have demonstrated that by fitting formant estimates to models, rapid estimation of frequency normalisation factors for speakers is possible. We have also shown that the normalisation is effective over all speech sounds, including unvoiced phonemes. Estimation of the normalisation factor requires only a limited amount of data and may then be applied to other utterances from the same speaker. This method is faster than techniques which use an exhaustive search to estimate the normalisation. Use of normalised data led to small improvements in recognition accuracy on the TIMIT database. Some interesting ways in which the technique might be improved are:

- use of a more accurate model for the formant distributions (e.g. Gaussian mixture)
- estimation of normalisation factors from unlabeled data

REFERENCES

[1] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *Proceedings of ICASSP 96, Atlanta*, pages 346 – 348, May 1996.

[2] L. Lee and R. C. Rose. Speaker normalization using efficient frequency warping procedures. In *Proceedings ICASSP 96, Atlanta*, pages 353 – 356, May 1996.

[3] W Press, B Flannery, S Teukolsky, and W Vetterling. *Numerical Recipies In C*, chapter 9. Cambridge University Press, 1988.

[4] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin. Speaker normalization on conversational telephone speech. In *Proceedings ICASSP 96, Atlanta*, pages 339–341, May 1996.

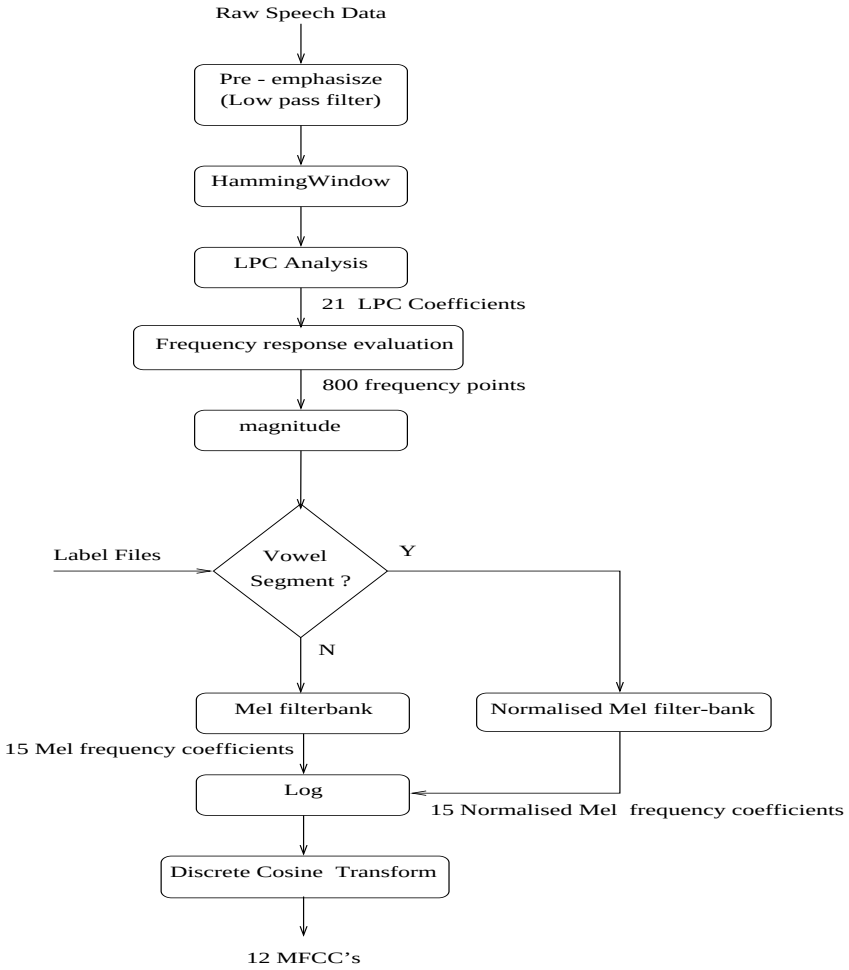

Figure 4. Vowel Normalisation

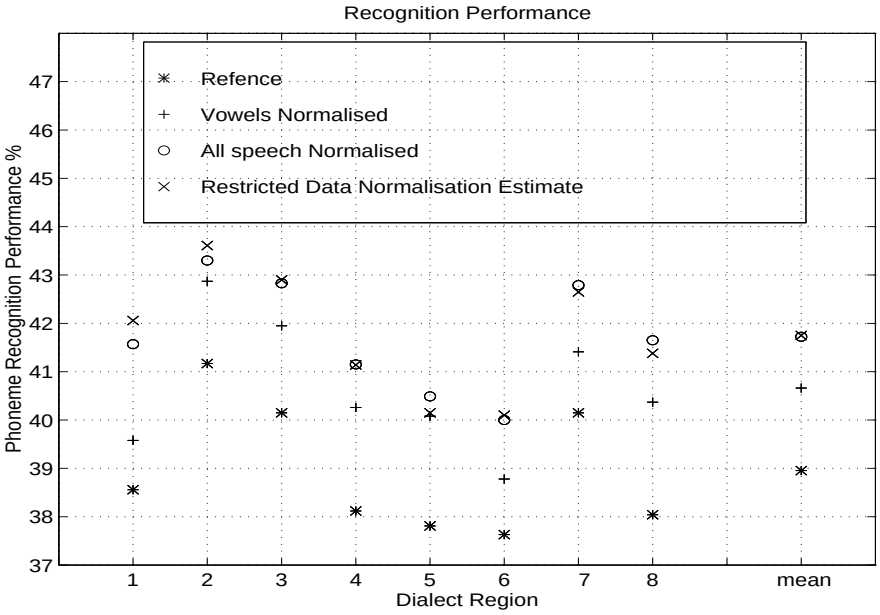

Figure 5. Recognition Test Results