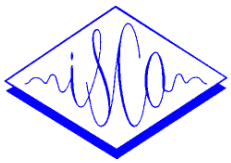

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

SPEAKER NORMALIZATION IN THE MFCC DOMAIN

Stephen Cox

School of Information Systems, University of East Anglia, Norwich NR4 7TJ, U.K.
sjc@sys.uea.ac.uk

ABSTRACT

It has been shown in several recent publications that application of vocal tract normalization (VTN) is a successful method for improving the accuracy of speaker independent recognisers. We argue that VTN can be implemented in the filterbank domain and propose a model to achieve this. We show how the model can be implemented directly in the MFCC domain, where it may be viewed as a constrained version of maximum likelihood linear regression (MLLR). The parameter estimates produced by the model are in accord with our ideas about how it should operate to perform VTN. Recognition results on a phoneme recognition task are presented which show a small improvement in accuracy.

1. INTRODUCTION

The variability in acoustic signals which are labelled as representing the same sound class comes from several sources. Some comes from physiological causes (e.g. vocal tract shape) and some from psychological (e.g. accent, mood, speaking rate), but any variation has an adverse effect on the performance of a speaker-independent (SI) speech recogniser.

One important source of variation is that due to the shape of a speaker's vocal tract. The length of the vocal tract can vary by as much as 25% between speakers and the effect of this is to cause differences in the short term spectrum of the same sound uttered by different speakers. Previous work [7], [9], [11], [4] has shown that the application of vocal tract normalization (VTN) to the utterances of a speaker may reduce inter-speaker variability caused by differing vocal tract lengths, hence reducing model variances (if used in training) and increasing recognition performance (if used in testing) when used in an SI speech recognition system.

A simple model of the vocal tract length predicts that if a speaker's vocal tract is $1/a$ times the average length, the spectrum of sounds produced will be scaled by a relative to an average spectrum. One technique for estimation of a is to use the positions of formants in the speaker's speech as "markers" and scale according to how far these are shifted relative to average [9], [4]. However, estimation of formant frequencies is usually undesirable. A more robust approach is to estimate a directly by an exhaustive search of a discrete set of normalization factors applied either to the waveform [1], or to estimates of the short-term spectrum of the signal [7]. In the latter case, a typical approach is to process the magnitude of the short-term FFT of the signal using filterbanks with different centre frequencies and bandwidths which simulate the effect of a scaling of the frequency axis. This is unattractive for the following reasons:

1. The search is typically carried out by processing the short-term magnitude FFT of the signal. This means that the FFT must be computed from the waveform and the technique cannot be applied if the signal is available only in a parameterized form. In addition, the FFT domain has high dimensionality so that this operation is computationally intensive.
2. There is some loss of accuracy because searching is only practically possible on a finite number of normalization factors.

If the short term spectrum of the speech signal is represented on a logarithmic frequency scale rather than on a linear scale, a scaling in frequency is manifest as a shift. Some evidence that this effect occurs in speech is presented in [2]. Here, it is shown that, if spectra are displayed on a Bark scale (which is approximately logarithmic), the spectra of some vowels spoken by a male can be quite accurately modelled by shifting down in frequency the corresponding spectra spoken by a female. A commonly used frequency scale for filterbank processing is the mel scale which is close in definition to the Bark scale [10]. In a mel-based filterbank, the centre frequencies of the filterbank are linearly spaced up to 1 kHz and logarithmically thereafter.

These observations suggest the model of speaker normalization proposed in this paper. The model is formulated as a shifting in the filterbank domain. Since the outputs from a mel filterbank are universally cosine-transformed to produce mel-frequency cepstral coefficients (MFCC's), we show how the model can be implemented directly in the MFCC domain. This model is not new, having been first proposed in [6] and previously investigated in [3]. However, it had previously not been viewed as a VTN technique and in this paper, we present evidence that it does indeed perform VTN. This is significant because it has been established that VTN can be succesfully combined with other speaker adaptation techniques (such as MLLR, [11]). We have also extended the model to enable it to have a variable number of shifts throughout the spectrum and extended the application of the technique to enable estimation to be done directly in the MFCC domain. The effectiveness of the model has been evaluated on a phoneme recognition task.

2. THE MODEL

We assume that, under the hidden Markov model (HMM) paradigm, a filterbank feature-vector $\boldsymbol{x}$ from a speaker was generated by a unimodal distribution D_i that is used to model sounds from a large number of speakers, and we designate this vector by $\boldsymbol{x}_i$. The vector consists of N components which are the outputs of N filterbank channels. Component k of $\boldsymbol{x}_i = x_i(k)$ is modelled as follows:

$$x(k) = \alpha\mu_i(k-1) + \beta\mu_i(k) + \gamma\mu_i(k+1) + \delta(k) \quad k = 1, 2, \ldots, N \tag{1}$$

where $\mu_i(k)$ is the value of the k'th channel of the mean of the distribution D_i and α, β, γ and $\boldsymbol{\delta}$ are the parameters of the model. (N.B. $\mu_i(0) = 0 = \mu_i(N+1)$). This model serves as a way of shifting the means relative to the data. If $\alpha = 1.0, \beta = 0.0, \gamma = 0.0$, the means are shifted upwards by one whole channel. Conversely, the values $\alpha = 0.0, \beta = 0.0, \gamma = 1.0$ give a downward shift of one channel. We assume that the shift is invariant for each sound and, in the model of equation 1, invariant over the spectrum (a model which has different shifts in different sections of the spectrum is introduced in section 4.2). The $\boldsymbol{\delta}$ term is required to normalize any offset between the feature vectors and the means. However, if cepstral mean normalization is used, $\boldsymbol{\delta}$ will be relatively small.

Shifting can be accomplished directly in the MFCC domain. The model of equation 1 can be represented in vector-matrix form as follows:

$$\boldsymbol{x}_i = A\boldsymbol{\mu}_i + \boldsymbol{\delta}. \tag{2}$$

where A is the tri-diagonal matrix:

$$A = \begin{pmatrix} \beta & \gamma & & & & & \\ \alpha & \beta & \gamma & & & & \\ & \alpha & \beta & \gamma & & O & \\ & & . & . & . & & \\ & & & . & . & . & \\ & O & & & \alpha & \beta & \gamma \\ & & & & & \alpha & \beta \end{pmatrix}$$

MFCC's are generated by applying an $M \times N$ cosine transformation matrix Φ to the $\boldsymbol{x}_i$, where M is the number of MFCC's used ($M \leq N$). The elements of Φ are

$$\Phi_{ij} = \sqrt{\frac{2}{N}} \cos\left(\frac{i(j-\frac{1}{2})\pi}{N}\right) \quad i = 1, \ldots, M \quad j = 1, \ldots, N. \tag{3}$$

Multiplying both sides of equation 2 by Φ gives

$$\Phi\boldsymbol{x}_i = \hat{\boldsymbol{x}}_i = \Phi A\boldsymbol{\mu}_i + \Phi\boldsymbol{\delta}, \tag{4}$$

where a circumflex denotes a cosine-transformed vector. Because we wish to work with MFCC data and means, we re-write equation 4 in terms of $\hat{\boldsymbol{\mu}}_i = \Phi\boldsymbol{\mu}_i$ as follows:

$$\begin{aligned} \hat{\boldsymbol{x}}_i &= \Phi A(\Phi^{-1}\hat{\boldsymbol{\mu}}_i) + \hat{\boldsymbol{\delta}} \\ &= \Phi A\Phi^T\hat{\boldsymbol{\mu}}_i + \hat{\boldsymbol{\delta}} \end{aligned} \tag{5}$$

If A is then decomposed into $A_1 + A_2 + A_3$, where A_1 consists of the α terms only, A_2 the β terms and A_3 the γ terms, the term $\Phi A\Phi^T$ in equation 5 can be written as

$$\Phi A\Phi^T = \alpha\Phi^1 + \beta I + \gamma\Phi^3, \tag{6}$$

where I is the identity matrix and the terms of the matrices Φ^1 and Φ^3 are respectively:

$$\Phi^1_{i,j} = \sum_{k=1}^{M-1} \Phi_{i,k+1}\Phi_{j,k} \tag{7}$$

$$\Phi^3_{i,j} = \sum_{k=1}^{M-1} \Phi_{i,k}\Phi_{j,k+1} \tag{8}$$

where the $\Phi_{i,j}$ are the coefficients of the cosine transformation matrix defined in equation 3. Hence the equation for the model in the MFCC domain is

$$\hat{\boldsymbol{x}}_i = (\alpha\Phi^1 + \beta I + \gamma\Phi^3)\hat{\boldsymbol{\mu}}_i + \hat{\boldsymbol{\delta}}. \tag{9}$$

The normalization parameters, $\boldsymbol{q} = \{\alpha, \beta, \gamma, \boldsymbol{\delta}\}$ are estimated by choosing parameters that maximise the likelihood of the total data X given the normalized models i.e.

$$\boldsymbol{q}^* = \arg\max_{\boldsymbol{q}} \Pr(X|M^*) = \prod_{j=1}^{X} \Pr(\hat{\boldsymbol{x}}^j_{d(j)}|M^*_{d(j)}), \tag{10}$$

where $d(j)$ is the index of the distribution of the j'th data example and M^* is the set of speech models whose means have been normalized according to

$$\hat{\boldsymbol{\mu}}^*_i = (\alpha\Phi^1 + \beta I + \gamma\Phi^3)\hat{\boldsymbol{\mu}}_i + \hat{\boldsymbol{\delta}}. \tag{11}$$

If we denote $\alpha\Phi^1 + \beta I + \gamma\Phi^3 = \Psi$, the transformation of the means is $\hat{\boldsymbol{\mu}}^*_i = \Psi\hat{\boldsymbol{\mu}}_i + \hat{\boldsymbol{\delta}}$ which has the same form as the linear transformation used in MLLR. However, unlike the MLLR transformation, this transformation is a highly constrained one designed to implement a particular model.

3. DATA AND SPEECH MODELS

The model was tested on a phoneme recognition task. The training set of the WSJCAM0 database [5] was used to train a set of 45 speaker independent (SI) hidden Markov models representing 44 phonemes and a silence model. Each model consisted of three states, plus an entry and exit state. A standard no-skip topology was used . Each state had a single Gaussian distribution with a diagonal covariance matrix associated with it. We used a unimodal distribution in these experiments because it was not clear what the effect of using a multimodal distribution (i.e. a Gaussian mixture) would be. To train the model, each frame is mapped to the mean of the distribution with which it is associated. The model is based on the premise that, on average, an individual speaker's frames are shifted up or down in frequency relative to the mean by a certain amount. However, the means of the components of a Gaussian mixture are distributed across the parameter space used by many speakers, and the position of a speaker's frame relative to the mean of the component most likely to have generated the frame is rather arbitrary. Hence for present purposes, we have sacrificed some baseline accuracy for the sake of gaining insight into how the model performs.

For parameter estimation and testing, we used the development set of WSJCAM0. This consists of about 100 sentences spoken by each of 20 speakers, 10 males and 10 females. Three sentences from each speaker were set aside for normalization use and all sentences were used for testing. The total number of sentences for testing was 1928, containing a total of 131663 phones.

The data were parameterized at 100 frames/second with a 26 channel filterbank, the outputs of which were transformed into 12 MFCC's. Energy, delta and acceleration coefficients were appended and cepstral mean normalization was used over each utterance. All processing described in this section was done using unmodified routines from the HTK (v2.1) software. The phone accuracy of the test set data when tested with the SI models was 42.8%.

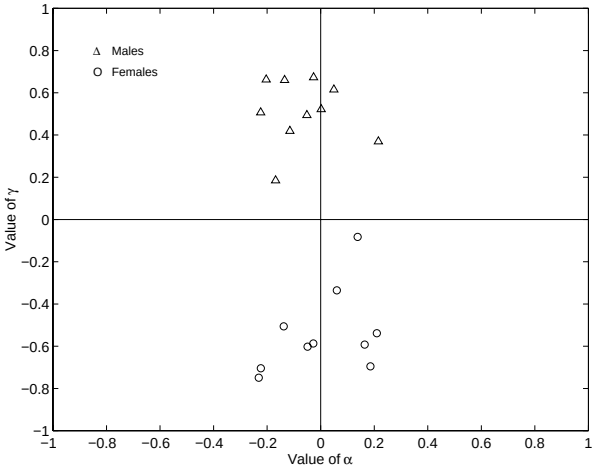

Figure 1: Single shift values for the 20 speakers

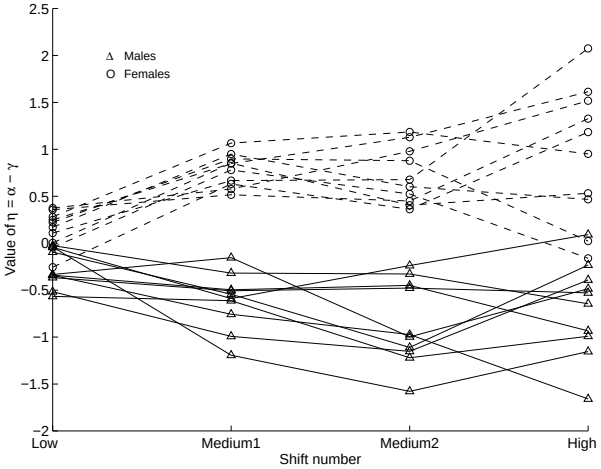

Figure 2: Magnitude of shift in four frequency regions for the 20 speakers

4. EXPERIMENTAL PROCEDURE

4.1. Single shift

The model requires that each speech frame is mapped to the mean of the distribution which "generated" the frame. These mappings were estimated by force-aligning the adaptation set utterances to the appropriate sequence of SI models as determined by the correct phonemic transcription of the utterance. Hence this is a supervised technique—however, a preliminary investigation in [3] developed an unsupervised version which displayed little decrease in accuracy over the supervised technique. The values of α, β, γ and $\boldsymbol{\delta}$ were then estimated using a quasi-Newton search method which uses the Broyden-Fletcher-Goldfarb-Shanno (BFGS) technique for updating the approximation of the Hessian matrix[1]. The estimated parameters were then used to adapt the means of the HMM's according to equation 11. The test-set data was then recognised using the normalized HMM's.

Figure 1 shows the values of α and γ estimated for each of the 20 speakers in the test-set. These values are in accord with intuition, as all male speakers have a positive value for γ and most of them a negative value for α: the net effect of these values is to shift the spectra of the SI means downwards. By contrast, all female speakers have a negative value for γ and most a positive value for α, which shifts the spectra of the SI means upwards.

4.2. Multiple shifts

The model of the vocal tract as a simple tube, which predicts a single shift over the entire spectrum, is a very over-simplified one. An obvious refinement is to allow different shifts in different regions of the spectrum, a technique which has also been proposed in [12]. The model of equation 1 can be extended to model multiple shifts in the spectrum i.e.

$$x(k) = \begin{cases} \alpha_1\mu_i(k-1) + \beta_1\mu_i(k) + \gamma_1\mu_i(k+1) + \delta(k) & k \le N_1 \\ \alpha_2\mu_i(k-1) + \beta_2\mu_i(k) + \gamma_2\mu_i(k+1) + \delta(k) & N_1+1 \le k \le N_2 \\ \text{etc.} \end{cases} \tag{12}$$

[1]The MATLAB routine `fminu`.

Equation 12 can be modelled in the MFCC domain after appropriate adjustments to equation 9. We experimented with multiple shifts by dividing the filterbank into 2,3,4,8 and 12 sections with approximately equal numbers of filters in each section. We also experimented with using a different shift for each filter (i.e. 26 shifts). A rough "magnitude" for a shift can be defined as $\eta = \alpha - \gamma$: η is positive for an upward shift in frequency and negative for a downward shift. Figure 2 shows the value of η when four separate shifts were used and confirms that the optimum shift is not constant over the spectrum. All speakers have a low value of shift (less than half a channel) in the lowest band (Low) and most a higher value in the next band (Medium1). In the two highest bands (Medium2 and High), the amount of shift is very variable.

4.3. Normalizing the data rather than the models

In some scenarios, it is more convenient to adapt the data rather than the HMM means. *A priori*, it might be thought that equation 1 could simply be reversed to model the data in terms of the means, which would then lead to the following expression for adapting the data:

$$\hat{\boldsymbol{x}}_i^* = (\alpha\Phi^1 + \beta I + \gamma\Phi^3)\hat{\boldsymbol{x}}_i + \hat{\boldsymbol{\delta}}. \tag{13}$$

However, when this model was implemented, it was found that although the likelihood of the adapted data given the SI models was much higher than the likelihood of the original data given the adapted models, the recognition accuracy fell to about one-third of the baseline accuracy. Examination of the normalization parameters showed that the values of α, β and γ were all in the region 0.2–0.4 for most speakers. This suggested that the transformation matrix A was smoothing the filterbank data rather than shifting it, which was confirmed by observing that the values of the high-order cepstrum coefficients were near zero in the adapted data. The adapted data fitted the mapped means much more tightly than the original data, but the lack of detail in the spectra of the data seemed to make it fit well to several other means, with the result that the recognition performance dropped. By contrast, if the means are adapted to the data, the transformation matrix cannot smooth the means to achieve the

Baseline accuracy (SI models) = **42.8%**

	No of adaptation utterances		
No of shifts	1	2	3
1	43.6	43.7	43.7
2	44.0	44.1	44.1
3	44.3	44.4	44.4
4	44.3	44.5	44.5
8	44.7	44.8	44.7
12	44.5	44.9	-
26	44.7	45.2	-

Table 1: Phoneme accuracy (%) using multiple shifts with different numbers of adaptation utterances

best fit to the data as the means are already smoothed versions of the data: hence it shifts the means, as intended.
If it is required to adapt the data rather than the means, one possibility is to estimate the normalization parameters by fitting the means to the data but to then apply the inverse transformation to the data rather than the forward transformation to the means. This raises the question of whether the matrix $(\alpha\Phi^1+\beta I+\gamma\Phi^3)$ is invertible, an issue currently under investigation.

5. RESULTS

Table 1 presents the results on the phoneme recognition task. The normalization technique shows a small improvement over the SI baseline which increases as the number of shifts is increased, with the highest accuracy being achieved when a different shift for each channel is used. Generally, the technique reaches optimum performance after two adaptation utterances. Note that a 1% increase in performance is equivalent to 1300 extra correct phonemes in the test set, so the increases in accuracy shown here are statistically highly significant. We are currently evaluating the performance of the model on digit string data supplied by AT&T which will enable us to compare the technique with the approach developed in [7].

6. DISCUSSION

We have introduced a model of speaker normalization that can be applied directly in the MFCC domain and which has some practical advantages over techniques which work in the waveform or FFT domain. Examination of the parameters estimated by the model suggest that the technique can be regarded as a form of vocal tract normalization. The model is essentially a constrained transformation applied in the MFCC domain, and as such could be regarded as being located somewhere between maximum likelihood linear regression (MLLR) [8] and the frequency-warping techniques currently used. The relationship of the constrained transformation proposed here to the general transformation used in MLLR is an interesting one which we intend to investigate further. If shifting in the filterbank domain is the major "distortion" between a speaker's data and the SI model, it is possible that MLLR is also essentially performing the kind of shift modelled here. If this is the case, the model proposed here should be more efficient, as the terms in the matrices Φ^1 and Φ^3 (which are due to implementing a shift in the filterbank domain in the MFCC domain) are defined explicitly, whereas they must be estimated in MLLR. Our immediate goals are to compare this technique in terms of both accuracy and speed of implementation with a frequency warping approach and to extend it to work with mixture distributions.

7. REFERENCES

[1] A. Andreou, T. Kamm, and J. Cohen. Experiments in vocal tract normalization. In *Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[2] R.A.W. Bladon, C.G. Henton, and J.B. Pickering. Towards an auditory theory of speaker normalization. *Language and Communication*, 4(1):59–69, 1984.

[3] S.J. Cox and J.S. Bridle. Simultaneous speaker normalisation and utterance labelling using Bayesian/neural-net techniques. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, pages 161–165, April 1990.

[4] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, volume 1, pages 346–349, 1996.

[5] J Fransen et al. WSJCAM0 corpus and recording description. Technical Report CUED/F-INFENG/TR.192, Cambridge University Engineering Department, September 1994.

[6] M.J. Hunt. Speaker adaptation for word based speech recognition systems (abstract only). *J. Acoust. Soc. Am.*, 69:S41–S42, 1981.

[7] L. Lee and R.C. Rose. A frequency warping approach to speaker normalization. *IEEE Transactions on Speech and Audio Processing*, 6(1):49–60, January 1998.

[8] C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 9(2):171–85, 1995.

[9] M Lincoln, S.J. Cox, and S.P.A. Ringland. A fast method of speaker normalisation using formant estimation. *Proc. 5th European Conference on Speech Communication and Technology*, pages 2095–2098, September 1997.

[10] D O'Shaughnessy. *Speech Communication—Human and Machine*. Addison-Wesley, 1987.

[11] D. Pye and P.C. Woodland. Experiments in speaker normalisation and adaptation for large vocabulary speech recognition. In *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing*, pages 1047–1050, May 1997.

[12] S. Umesh, L. Cohen, and D. Nelson. Frequency warping and speaker normalization. In *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing*, pages 983–986, April 1997.