GENDER EFFECTS IN SPEAKER RECOGNITION

J.S. Mason & J. Thompson
Department of Electrical & Electronic Engineering,
University College of Swansea, SWANSEA, SA2 8PP, UK

ABSTRACT

In this paper we examine the properties of male and female speech and its effect on speaker recognition. In particular we compare the recognition performance of males and females, and find females perform consistently worse than males, when using MFCCs, by a factor of about 5%. To rectify this, we explore how modifying the frequency warping functions (mel and bark) can improve the performance of both genders. We find the mel scale optimum for males, and the linear frequency scale (i. e. no warping) is optimal for females. Also we show the influence of pitch in the mel cepstrum and LDA transform using simulated data.

1 Introduction

Speaker recognition relies on the differences in the acoustical properties of the speech waveform between humans. These differences can be attributed to two sources anatomy, and pronunciation. Anatomical differences contribute important inter-speaker variations in speech. In *speaker* recognition these differences are the most beneficial components within any feature set: they are more likely to be consistent over time and they are not easily mimiced. In contrast, in speaker independent *speech* recognition such differences tend to be viewed as just another source of variation to be overcome by robust features or modelling. So, what is primary discrimination material in one case, is of nuisance value in the other. It is somewhat surprising therefore that similar front-end processing can to be adopted in both applications: for example mel-scaled cepstra coefficients (MFCC) and their *Delta* forms are popular in both areas. Analysis of the anatomical differences between genders through simulation, and experimentation, leads to some revelations concerning the performance of speaker recognition systems.

In previous work we have demonstrated the importance of higher order cepstra [1], and the benefit of frequency resolution (more filters) in the log spectral estimates [2], both in the context of *speaker* recognition. These enhancements might be specific to the particular task, the case as yet is unproven.

In this paper we focus on differences in male and female speech, as it pertains to both *speech* and *speaker* recognition. Like others, we have observed that discriminating between female speakers in measurably more difficult than discriminating between males (see Figure 1). In an attempt to improve this performance, we review findings of others and investigate high-order cepstra, their weightings and representation in terms of vocal tract and excitation components.

2 The Problem

Female speakers (and children) are more difficult to discriminate than males. The primary reason might well be an inherent one that anatomical variations are less. Experimental results are shown in Figure 1: 4 profiles of speaker identification (SI) errors against cepstral order, MFCC-5 up to MFCC-17, using an alphabet-independent, codebook classifier and inverse variance weighting. For the spectra, we use standard triangular weightings on a mel scale [3], but rather than 24 filters (0 to 5kHz), we use 32 filters following the findings of [2].

Each profile represents experiments on 20 speakers, 2 sets of males and 2 sets of females (80 different speakers in all). The discrepancies between male and female sets are large and consistent. Our goal is to improve feature performance, particularly in the case of the female sets.

3 High Weightings and Higher Order

It is clear from Figure 1 that performance continues to improve, albeit slowly, with orders above the more normal 10 or 12 used in *speech* recognition. This is in agreement with findings in [1], using the LPC-based PLP feature. However, a word of caution is needed. Figure 2 shows cepstral variations against *simulated* pitch pulses up to a frequency of 500 Hz. It is clear that excitation

ASR error (%)		
	Male	Female
IMELDA	1.3	3.7
mel-cep	5.3	5.3
mel-cep var	4.2	15.2
mel-cep quef	6.2	22.7

Table 1: Male and female *speech* recognition errors (%) for different weightings: quef - quefrency, index weighting; var - inverse variance, taken from Hunt [5]

components break through into the higher order cepstra in a significant way, and while pitch can be useful as a distinguishing feature, it is known to lack consistency in field-trial situations [4].

Hunt's findings [5] in speaker independent *speech* recognition, separating the results into male and female is of interest here. A summary taken from [5] is given in Table 1, and shows that while inverse variance and quefrency weightings give good results for males speakers, both are significant over weightings for females, with a resultant notable decrease in performance. This is very different to our experiments on *speaker* recognition, where higher cepstral weightings prove beneficial in both male and female cases. The best results Hunt obtained using IMELDA [6].

IMELDA is a system which performs a data-dependent transform, from linear discriminant analysis (LDA), upon filter bank outputs. In the absence of its cosine transform in IMELDA, it is possible that more excitation components are reflected in the resultant features. Of course if the training data is truly representative, the this is perfectly acceptable. However, as we have mentioned, pitch can be somewhat fickle. We test the homomorphic filtering action of IMELDA, again using the artificial data, the results shown in Figure 3. Indeed in comparison with MFCC it can be seen that the lower order IMELDA coefficients are influenced by pitch, and hence our preference for combining LDA and the cosine transform [7]

4 Frequency Warping

The effects of frequency warping are demonstrated in reports by Gu [9] and Shikano and Itakura in [10]. Both investigate the effects of changing the warping functions, Gu raising the linear-to-Bark conversion to a power, α

$$w_{warp} = f(w_{lin})^{\alpha} \quad (1)$$

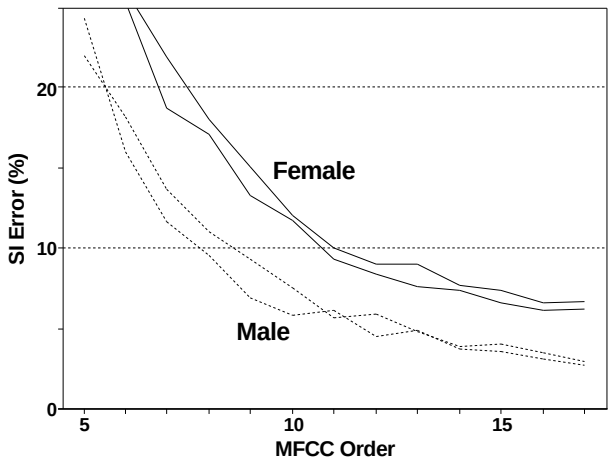

Figure 1: Speaker identification errors (%) versus cepstral order, for 4 sets of speakers: 2 male and 2 female; the latter groups give consistently more errors

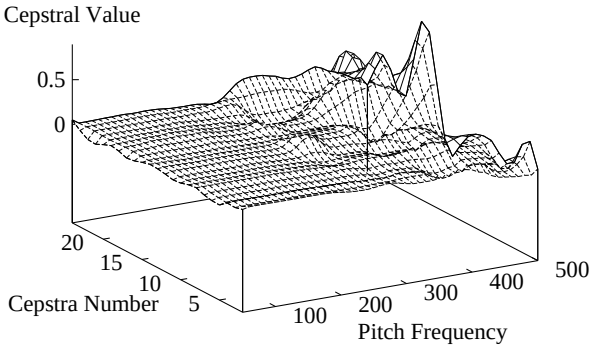

Figure 2: Cepstra values: 'quefrency' versus pitch frequency for simulated pitch pulses

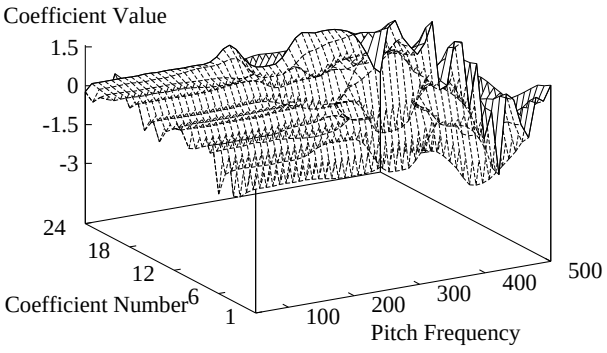

Figure 3: LDA values: LDA coefficients versus pitch frequency for simulated pitch pulses

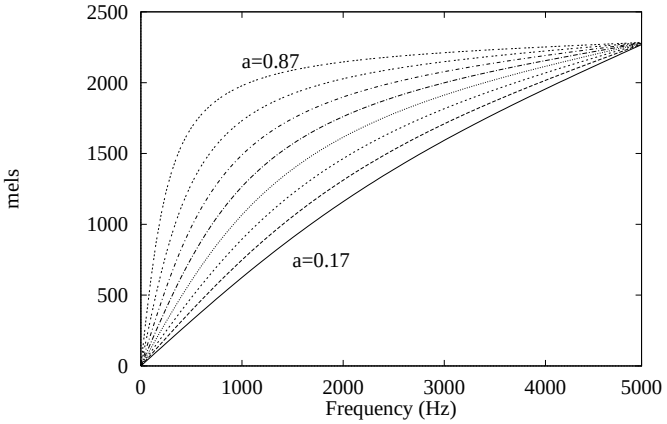

Figure 4: Warping functions for different values of 'a' in the bilinear transform; filter number versus linear frequency (0 to 5kHz.)

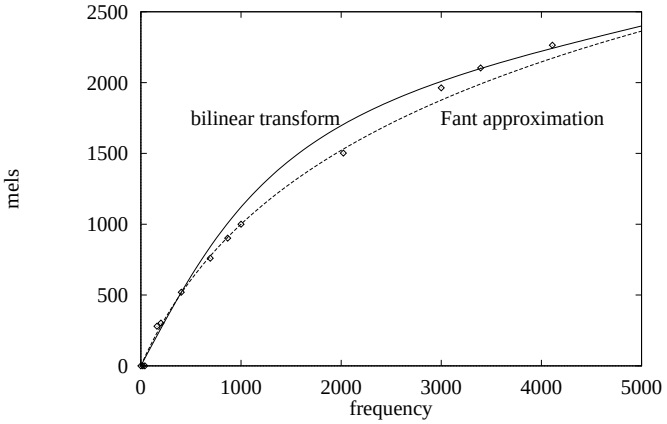

Figure 5: Accuracy of Fant and bilinear approximations of the mel scale, when compared with Stevens original [8].

while Shikano and Itakura use the bilinear transform approximation of the mel scale

$$Z^{-1}_{warp} = \frac{z^{-1} - a}{1 - az^{-1}} \qquad -1 < a < 1 \tag{2}$$

see Figure 4, showing profiles for a=0.17 to 0.87 in 0.1 steps. A value of a=0.47 is very close to the popular Fant approximation form of the mel scale [11]:

$$log_{10}(1 + f/700) \times 2595 \tag{3}$$

A comparison of these two approximations is shown in Figure 5 and their accuracy compared with Steven's original measurements [8]. Interestingly, Gu shows a distinct optimum for female speakers of $\alpha = 1.2$, but for male speakers the normal value of $\alpha = 1.0$ is best, whereas Shikano shows significant variations of optimum warping across different speech classes. In both cases the above work relates to *speech* recognition. The effect of frequency warping in the context of *speaker* recognition performance, using the warping functions shown in Figure 4, agrees with the findings of Gu. In the case of females we find an optimum with a less severe warping function than imposed by the commonly used Fant approximation of the mel scale of Equation 3. For males, a higher degree of warping has been found to be optimum. These results are broadly agree with the findings of Gu [9].

5 Explanation of Findings

To explain these findings, two aspects of the cepstrum must be considered, firstly the formant structure and secondly the pitch. The cepstrum of typical voiced speech samples are shown in Figures 6 and 7. Figure 6 is obtained from a female speaker, while Figure 7 is from a male. There are two significant differences between these two cepstra: the pitch component amplitude in the female speech is significantly higher than pitch in the male speech; and the formant structure of the the female speech is concentrated below the quefrency index of 5ms, whereas the formant structure of the male speech is extends up to a quefrency index of 7.5ms, a difference of 50%. These two factors have considerable bearing upon the speaker recognition results. Firstly the mel scale frequency warping concentrates the positioning of filters below 1KHz. This is fine for male speech because the first few formants are concentrated in this region, and thus good recognition performance is achieved. These 'low' frequency formants map onto the quefrency scale of Figure 7 at points near or below 7.5mS. In female speech these first formants, are considerably higher in frequency (lower in quefrency). thus few mel scale filters concentrated in this region of the spectrum less resolution is obtained. When the mel scale is removed, the filters are evenly distributed, and therefore the formants of the female speech are detected with increased resolution. The effect of using the mel scale on the cepstrum is to concentrate female speech in quefrency region even lower than the 5mS shown in Figure 6. Also, with the mel scale the pitch of female speech, due to its relative high frequency, is not separated from the formants to a great extent, therefore by using say the first 14 cepstra in a recognition system, the pitch will also be present. This effect has also been noted by Thompson [12]. With no mel frequency warping, the pitch is separated to a greater extent from the formant structure, due to the even distribution of filters. This also accounts for the findings of Gu [9] and the different optimum warping functions for male and female speakers.

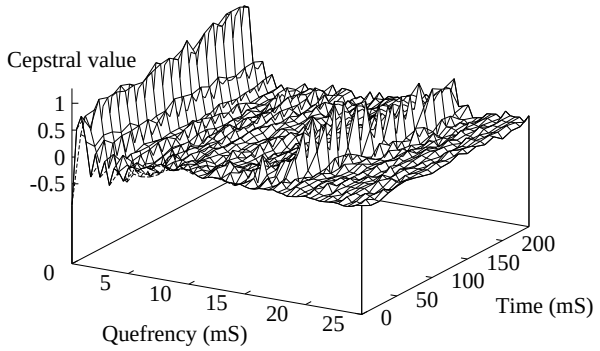

Figure 6: Cepstrum of female uttering /e/ with no frequency warping

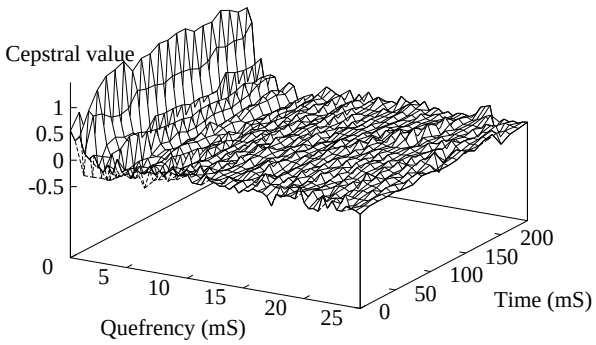

Figure 7: Cepstrum of male uttering /e/ with no frequency warping

6 Conclusions

- Anatomical characteristics of speakers to a large extent determine the performance of speaker recognition systems, with females performing consistently worse than males.
- The mel scale is almost optimal for males, but suboptimal for females in speaker recognition, because it quefrency quashes the cepstra containing the female formants.
- The influence of pitch on MFCCs is large for females, but almost non-existent for males.
- The separation between formants and pitch in the cepstrum decreases rapidly as pitch increases. With the mel scale this decrease becomes even more rapid.

References

[1] L. Xu, J. Oglesby, and J. S. Mason. The optimization of perceptually-based features for speaker identification. *Proc. ICASSP-89, Vol. 1*, pages 520–523, May 1989.

[2] L. Xu and J.S. Mason. Optimization of perceptually-based spectral transforms in speaker identification. *Proc. Eurospeech-91*, pages 439–442, 1991.

[3] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. ASSP-28*, pages 357–366, 1980.

[4] J. Eatock and J. S. Mason. Phoneme performance in speaker recognition. *Proc. ICSLP-92, Canada*, pages 1411–1415, 1992.

[5] M. J. Hunt and C. Lefebvre. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. *Proc. ICASSP-89*, pages 262–265, 1989.

[6] M. J. Hunt and C. Lefebvre. Speaker dependent and independent speech recognition experiements with an Auditory Model. *Proc. ICASSP-88*, pages 215–218, 1988.

[7] J. P. Openshaw, Z. P. Sun, and J. S. Mason. A comparison of feature performance under degraded speech in speaker recognition. *Proc. ESCA-92*, 1992.

[8] S. S. Stevens and J. Volkman. The relation of pitch frequency: a revised scale. *American Journal of Psychology*, 53:329, 1940.

[9] Y. Gu. Perceptually-based features in automatic speech recognition. *Ph.D. Thesis, University College Swansea*, 1990.

[10] S. Furui and M. M. Sondhi, editors. *Advances in Speech Signal Processing*. Marcel-Dekker, 1991.

[11] C. G. M. Fant. Acoustic description and classification of phonetic units. *Ericsson Technics*, 1, 1959.

[12] J. Thompson and J. S. Mason. Cepstral statistics within phonetic subgroups. *Proc. ICSP-93, Beijing*, pages 737–740, 1993.

Acknowledgement This work has been supported by Ensigma Ltd. of Chepstow, U.K.