Exploitation of spectral variance to improve robustness in speech recognition

H. Xu, Z.-H. Tan, P. Dalsgaard and B. Lindberg

With the aim of improving noise robustness of speech recognition an approach that exploits the variance information in spectral sub-bands is presented. The variance based features are used in combination with the normally used Mel-frequency cepstral coefficients (MFCC), and experimental results show that the combined features outperform MFCC alone, perceptual linear prediction features and entropy based features.

Introduction: The pre-processing in automatic speech recognition (ASR) systems extracts a set of speech features with the purposes of achieving high discrimination among recognition classes and maintaining the performance for noise corrupted speech signals.

The commonly used MFCC features [1] are calculated from Mel-filter outputs mainly reflecting information of sub-band spectral mean values. It is noted, however, that spectral mean values are highly dependent on the additive noise occurring in speech signals, and thus cause a drop in ASR performance. With the aim of counteracting this influence, entropy-based features [2] were proposed to convey the peak energy in each band as opposed to the mean values of the MFCC features. In [2] the entropy-based features were combined with perceptual linear prediction features (PLP) [3] also and showed a slight improvement over the PLP alone.

This Letter investigates the variances of the speech magnitude spectrum in Mel sub-bands resulting in variance-based MFCC (VMFCC) features. The variance focuses on the dynamically changing information and the VMFCC features are thus expected to be more noise robust.

The combination of MFCC and VMFCC is compared with MFCC alone, with PLP features in addition to entropy based features and shows better recognition performance.

MFCC and VMFCC: The MFCC features are calculated based on the magnitude spectrum in Mel-filter bands. As shown in Fig. 1, mean values of the Mel-filtered magnitude spectrum are calculated, and transformed by the logarithmic function and the discrete cosine transform together, resulting in the final cepstrum. However, the mean values are vulnerable to additive noise, which may lead to poorer ASR performance. In addition, the magnitude spectrum of a voiced speech segment is generally characterised by a number of peaks and valleys, and by only using the mean values in MFCC detailed information of the dynamical change in each band is lost.

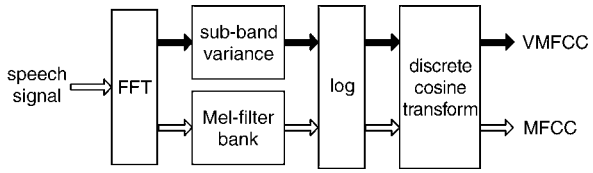

Fig. 1 *Feature extraction processes*

VMFCC features are introduced with the goal of maintaining ASR performance for corrupted speech signals. The calculation of the VMFCC features are conducted based on the variance of the unfiltered magnitude spectrum in each Mel sub-band as shown in Fig. 1. Compared to the MFCC features, the VMFCC features represent the dynamic variation within each band.

It is noted that calculation of the VMFCC removes the mean value of the combined speech and additive noise signal from the spectrum rendering the VMFCC features less sensitive to noise. For the extreme case with full-band white additive noise, the VMFCC features are – in a statistical sense – not influenced by noise.

Mel-grams for the MFCC (measuring the output of Mel-filter bank) are compared with Mel-grams for the VMFCC (measuring the output of sub-band variance) as shown in Fig. 2. It is observed that the VMFCC Mel-grams are less influenced by noise than the MFCC Mel-grams. However, in calculating the variances, the VMFCC features also 'filter out' the relatively flat speech valleys resulting in the loss of information here.

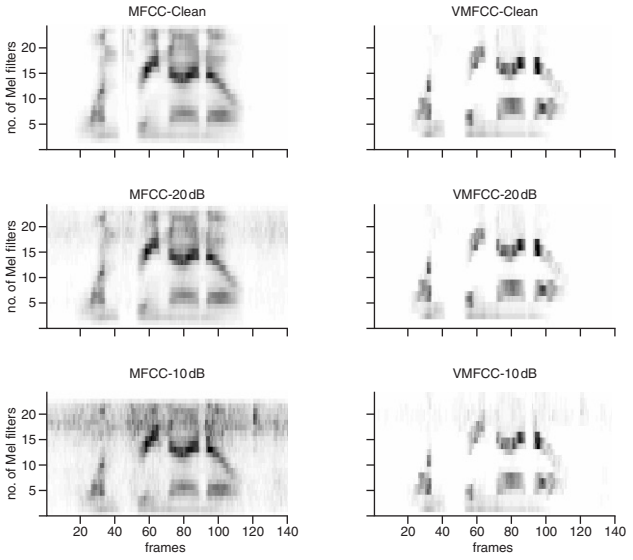

Fig. 2 *Comparisons of Mel-grams for MFCC and VMFCC for utterance 'one three nine oh' corrupted by 'subway' noise with different signal-to-noise ratios (SNRs)*

It may therefore be advantageous to combine the VMFCC with the MFCC features. In this Letter the combination is carried out – for simplicity – by appending the VMFCC to the original MFCC features. A reduction of the combined feature length can be achieved, for example, by applying LDA or multi-stream methods but experiments on this are not included in this Letter.

Experiments: A set of experiments have been conducted to test the robustness by measuring the ASR performance. The evaluations are based on the continuous English digits speech recognition task Aurora 2 [4], encompassing two training sets (clean training and multi-condition training) and three test sets. In the experiments, the clean speech material is used for training the hidden Markov model (HMM) models and test set A is selected as the only test set which includes the speech data contaminated with four types of additive noise (subway, babble, car and exhibition) with SNR ranging from 20 to 0 dB. Each HMM model has 16 states with three Gaussian mixtures per state whereas 'silence' is modelled with three states each with six Gaussian mixtures.

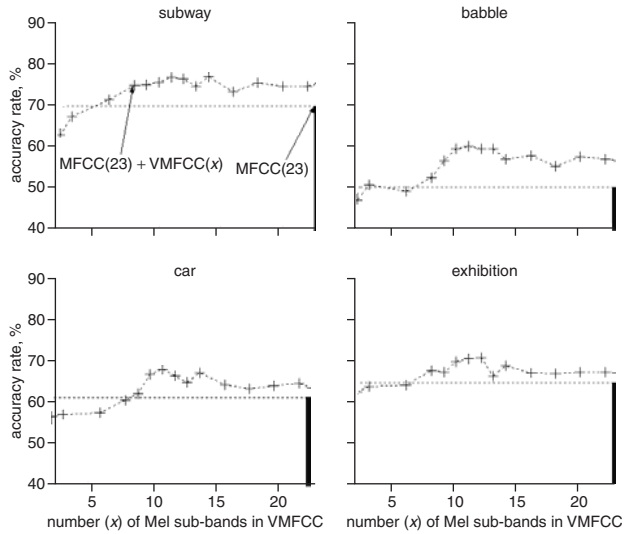

Fig. 3 *Recognition accuracy (%) (averaged over all SNRs) for MFCC with 23 sub-bands and VMFCC with variable number (denoted as x) of sub-bands*

In the first experiment, the MFCC features are calculated on the basis of 23 Mel-filters and encompass 12 cepstral components with the zeroth component excluded. The number of sub-bands for calculating the VMFCC features varies from 1 to 23. When the number is equal to or

larger than 13, VMFCC features include 12 cepstral components the same as the MFCC, and otherwise include all components except the zeroth component.

In all experiments the MFCC features, the VMFCC features, and logarithmic energy (LogE) are appended with their corresponding velocity and acceleration components.

Fig. 3 shows that the combined MFCC and VMFCC features start to outperform the purely MFCC features for the band number ranging from six to nine dependent on noise type and achieve the best results at 11 for all the tested noise types. The effect of the sub-band number on recognition performance may be explained as the need for a trade-off between the sub-band resolution and the robust variance estimation.

The results of the second set of experiments are shown in Fig. 4 comparing four different feature combinations. The PLP features are constructed from HTK [5] with 12 components whereas the 'Entropy + PLP' features are obtained as in [2]. The results show that the combined 'VMFCC(11) + MFCC(23)' features give the highest ASR performance over all the SNR values.

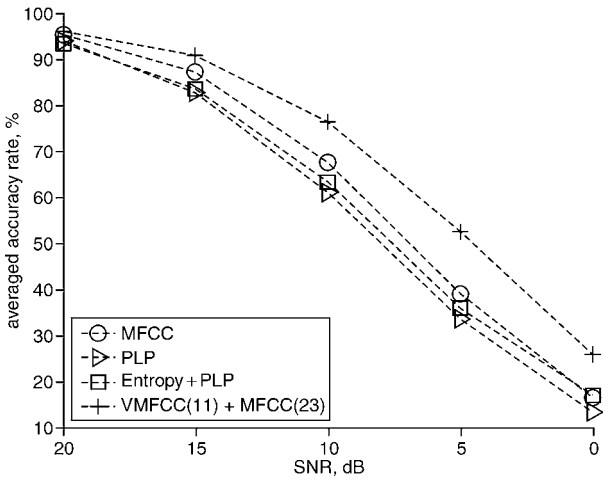

Fig. 4 *Recognition accuracy (%) (averaged over four noise types) against SNR for four feature combinations with VMFCC using eleven sub-bands*

Conclusion: In this Letter, spectral variance based features are introduced by calculating the variances of the unfiltered magnitude spectrum in each Mel-band. The variance features are used in combination with the normally used MFCC features with the aim of improving ASR noise robustness. Test results show that the proposed feature combination outperforms the MFCC, PLP and 'Entropy + PLP' based features justifying that sub-band variance includes relevant speech information and is helpful for noise robustness.

4 November 2005
Electronics Letters online no: 20063884
doi: 10.1049/el:20063884

H. Xu, Z.-H. Tan, P. Dalsgaard and B. Lindberg (*Department of Communication Technology, Aalborg University, 9220 Aalborg, Denmark*)

E-mail: hx@kom.aau.dk

References

1 Davis, S.B., and Mermelstein, P.: 'Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences', *IEEE Trans. Acoust. Speech Signal Process.*, 1980, **28**, (4), pp. 357–366
2 Misra, H., Ikbal, S., Sivadas, S., and Bourlard, H.: 'Multi-resolution spectral entropy feature for robust ASR'. 2005 IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP'05), Philadelphia, PA, USA, March 2005
3 Hermansky, H.: 'Perceptual linear predictive (PLP) analysis of speech', *J. Acoust. Soc. Am.*, 1990, **87**, (4), pp. 1738–1752
4 Hirsch, H.G., and Pearce, D.: 'The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions'. Proc. ISCA ITRW ASR2000, Paris, France, September 2000
5 Young, S.J.: 'HTK: hidden Markov model toolkit V3.2.1', *Reference Manual*, Cambridge University Speech Group, March 2004