UNSUPERVISED NEURAL-NETWORK BASED VOCAL TRACT LENGTH NORMALIZATION

Philip Harding

Nuance Communications Ltd. UK

philip.harding@nuance.com

ABSTRACT

In this paper an efficient, unsupervised, method of warp factor estimation for vocal tract length normalisation (VTLN) is proposed. VTLN is a method of feature-based speaker normalisation whereby the frequency spectrum is warped to produce speaker-independent spectral features that are invariant to vocal tract length. The degree to which the spectrum is warped is determined by the warping factor, and it is the estimation of this warping factor that is the focus of this paper.

The warping factor is typically obtained using a maximum likelihood-based technique that requires a state alignment for each utterance and a GMM acoustic model trained on warped features. The proposed method of warp factor estimation makes use of a small neural network, trained on un-warped features, to directly estimate the warping factor. Experimental results are presented where, unlike previously published methods of unsupervised warp factor estimation [1,2], the proposed method is shown to give equivalent performance to the typical supervised GMM-based method in terms of ASR accuracy at a significantly lower computational cost.

Index Terms: speech recognition, speaker normalisation, vocal tract length normalisation

1. References

[1] P. Cerva et al., "Using Unsupervised Feature-Based Speaker Adaptation for Improved Transcription of Spoken Archives.," in *INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association, August 27–31, Florence, Italy, Proceedings*, 2011, pp. 2565-2568

[2] Paczolay, Dénes, András Kocsor, and László Tóth, "Real-time Vocal Tract Length Normalization in a Phonological Awareness Teaching System.," in *Text, Speech and Dialogue 2003* – pp. 309-314 Springer Berlin/Heidelberg.