Studentská Vědecká Konference 2011

VTLN LINEAR TRANSFORMATION USING SUFFICIENT STATISTICS

Zbyněk ZAJÍC [1]

1 INTRODUCTION

The accuracy of an acoustics model in a speech recognition system depends beyond others on an actual test speaker. There are many sources of an inter-speaker variation, one of these is a vocal tract length of the speaker. The vocal tract significantly affects the position of formant frequencies. Vocal Tract Length Normalization (VTLN) (Zhan (1997)) is one of speaker adaptation methods of an acoustics model. VTLN transforms the frequency axis of the speaker to normalize the position of his formants. However, classic VTLN is very time consuming (especially the estimation of the warping parameter, more in Section 2.1) compared with VTLN as a linear transformation, see Section 3.

2 FREQUENCY WARPING FUNCTION

Transformation of a frequency axis is usually performed by nonlinear warping of the frequency scale. There are many warping functions $\tilde{\omega} = \mathcal{F}^{\alpha}(\omega)$, the simplest approach uses a linear warp of frequency ω, however, for complex warping the bilinear function is used:

$$\mathcal{F}^{\alpha}(\omega) = \omega + 2\arctan\left(\frac{(1-\alpha)\sin\omega}{1-(1-\alpha)\sin\omega}\right), \tag{1}$$

where α is a warping factor. Given a warping function, normalization can by implemented either by re-sampling and interpolating the spectrum or (for MFCCs – Mel-Frequency Cepstral Coefficients) by warping frequencies of the mel filter bank.

2.1 Estimation of the warping factor

If train utterances of the speaker s are $\boldsymbol{O}_s = \{\boldsymbol{O}_s^1, \ldots, \boldsymbol{O}_s^E\}$ and relevant transcriptions are $\boldsymbol{W}_s = \{\boldsymbol{W}_s^1, \ldots, \boldsymbol{W}_s^E\}$, then the α-warped utterances are $\boldsymbol{O}_s^{\alpha} = \{\boldsymbol{O}_s^{1\alpha}, \ldots, \boldsymbol{O}_s^{E\alpha}\}$. Finding an optimal warping factor α_s^* is an optimization process, α_s^* can be found by maximization of the likelihood of warped utterances $\boldsymbol{O}_s$ with respect to the acoustics model λ and transcriptions $\boldsymbol{W}_s$:

$$\alpha_s^* = \arg\max_{\alpha} P(\boldsymbol{O}_s^{\alpha}|\lambda, \boldsymbol{W}_s). \tag{2}$$

α_s^* is usually searched in the interval $\langle 0,88; 1,12\rangle$.

3 VTLN USING A LINEAR TRANSFORMATION (VTLN-LT)

In the previous section, the estimation of α_s^* required a new parametrization for each tested $\alpha \in \langle 0,88; 1,12\rangle$. In Umesh (2005) was introduced a new approach using only statistics of speaker adaptation data $\boldsymbol{O}_s$, VTLN is done by a linear transformation of

[1]Ing. Zbyněk Zajíc, Ph.D. student,University of West Bohemia in Pilsen, Faculty of Applied Sciences, Department of Cybernetics, Univerzitní 22, 306 14 Pilsen, e-mail: zzajic@kky.zcu.cz

cepstra $\boldsymbol{C}_s = \boldsymbol{O}_s/_{MFCC}$ instead of a nonlinear warping of frequencies of the mel filter bank.

3.1 Linear transformation

Non-warped cepstrum $C^{1,00} = DCT[\log(F_m S)]$ is given by a spectrum of the original signal S filtered by mel filter bank F_m, DCT stands for Discrete Cosine Transformation. Warped cepstrum $C^{\alpha} = DCT[\log(F_m^{\alpha} S)]$ differs in warping frequencies of the mel filter bank F_m^{α}. Relation between $C^{1,00}$ and C^{α}:

$$C^{\alpha} = DCT[\log(F_m^{\alpha}\{F_m^{-1} \exp DCT^{-1}(C^{1,00})\})] \tag{3}$$

can be written (after some approximation in Panchapagesan (2008)) as linear transformation W^{α}:

$$C^{\alpha} = (DCT\ DCT^{\alpha T})C^{1,00} = W^{\alpha}C^{1,00}. \tag{4}$$

When considering DCT as an unitar matrix II-type DCT, warped DCT matrix is given by a formula:

$$DCT^{\alpha} = \left[\cos(\pi n \mathcal{F}^{\alpha}(\frac{2m-1}{2M})\right]_{0 \leq n \leq N-1, 1 \leq m \leq M}, \tag{5}$$

where M is the number of mel filters and N is the number of cepstra.

3.2 Estimation of the warping factor from adaptation statistics

Because VTLN as linear transformation is in use, we can warp speaker adaptation statistics and rewrite the auxiliary function (2) into the form as in fMLLR approach (Povey (2006)). In praxis, it is possible to precompute matrices $W^{\alpha}|_{0,88 \leq \alpha \leq 1,12}$ (they depend only on M an N) and find the one which maximizes the auxiliary function for speaker adaptation data.

4 CONCLUSION

VTLN approach is based on the vector normalization. Warped vectors $\boldsymbol{O}_s^{\alpha}$ can not be recognized by an original acoustics model λ, instead of a new normalized model λ_c must be retrained on warped vectors. The model λ_c represents an average speaker with a normalized length of the vocal tract. Recognition with VTLN adaptation improves the accuracy about 1-2% absolutely. VTLN-LT can be used in the on-line adaptation as an unsupervised and very fast adaptation approach, which gives same results as VTLN.

REFERENCES

Zhan, P. and Westphal. M., 1997. Speaker Normalization Based On Frequency Warping in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 3. pp 1039 – 1042.

Umesh,S. and Zolnay, A. and Ney, H., 2005. Implementing Frequency-Warping and VTLN Through Linear Transformation of Conventional MFCC in *Interspeech*, pp 269 –272.

Panchapagesan,S. and Alwan, A., 2008. Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC in *Computer Speech and Language*, pp 42 – 64.

Povey, D., Saon, G., 2006. Feature and Model Space Speaker Adaptation with Full Covariance Gaussians. In: Interspeech, paper 2050-Tue2BuP.14.