SPEAKER NORMALIZATION FOR AUTOMATIC SPEECH RECOGNITION - AN ON-LINE APPROACH *

Ioannis Dologlou, Tom Claes°, Louis ten Bosch°, Dirk Van Compernolle°, Hugo Van Hamme°*
* Katholieke Universiteit Leuven - ESAT - PSI
° Lernout & Hauspie Speech Products, Belgium
Tel: +32 16 321827; fax: +32 16 321723
e-mail: dologlou@esat.kuleuven.ac.be

ABSTRACT

We propose a method to transform the on line speech signal so as to comply with the specifications of an HMM-based automatic speech recognizer. The spectrum of the input signal undergoes a vocal tract length (VTL) normalization based on differences of the average third formant F_3. The high frequency gap which is generated after scaling is estimated by means of an extrapolation scheme. Mel scale cepstral coefficients (MFCC) are used along with delta and delta^2-cepstra as well as delta and delta^2 energy. The method has been tested on the TI digits database which contains adult and kids speech providing substantial gains with respect to non normalized speech.

1 INTRODUCTION

This paper proposes a method to transform the input speech signal enabling the efficient use of already trained HMM's for speakers with different vocal tract characteristics.

It is known from literature [1][2][3] that the spectral properties of male, female and child speech differ in a number of ways. One prominent difference is due to the difference between their average vocal tract length (VTL). In fact, the VTL of females is about 10% shorter compared to the VTL of males. The VTL of children is even shorter (up to 10 %) than that of the females. According to the linear acoustic theory of speech production, this directly implies that all the formants in male speech undergo a (fixed, VTL-dependent) scaling towards the high end of the spectrum. Consequently, one has to warp the children spectra towards the lower end in order to match it with the adults (male) speech. In this paper, the problem of how to do so is addressed.

Another important issue which is also discussed here is related to the estimation of the VTL from a given speech signal. This is used subsequently to determine the frequency warping factor. It is known that the VTL is related to the position of formants and in particular to the position of the third formant (F_3). As opposed to the values of F_1 and F_2, the value of F_3 is less influenced by the vowel under consideration while its detection from the signal itself remains quite reliable. For that reason the estimation of both the VTL and the warping factor are based on the F_3 values [11].

The feature vector, which is computed from a speech frame, consists of the mel scale cepstral coefficients (MFCC) (12 cepstra) along with 12 delta-cepstra, 12 delta^2-cepstra and delta-log(Energy), delta^2-log(Energy) , i.e. 38 parameters in total.

The MFCC parameters [8] are calculated from the power spectrum using a simulated filterbank with triangular filters. The filter centers are linearly spaced below 1kHz and logarithmically above that. The mel cepstrum is then calculated by taking the logarithm and the inverse cosine transform (IDCT) fig. 1.

First the input signal spectrum is scaled downwards by a factor which is determined from the ratio of the F_3 frequencies of the two groups. Thereafter the generated high frequency gap in the power spectrum is estimated by means of interpolation and/or extrapolation techniques. To evaluate the proposed methodologies, experiments for continuous word speech recognition have been carried out using the digit-strings of the TI-digits database.

*This research work was carried out at the ESAT laboratory of the Katholieke Universiteit Leuven, supported by the *KIDS* Research Contract nr. ITA/950226 from the IWT.

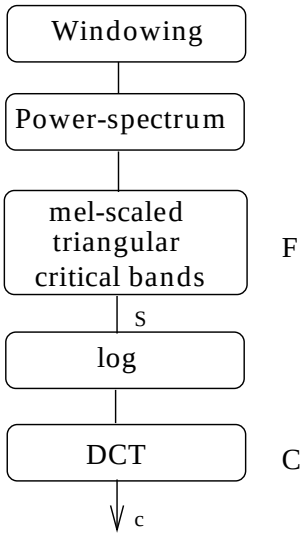

Figure 1: Calculation scheme for MFCC parameters

2 WARPING TECHNIQUE FOR VOCAL TRACT LENGTH NORMALIZATION

The difference in average VTL for Male, Female and Kids is an important problem for speech recognition with children compared to speech recognition with adults. The VTL is related to the i-th average formant with (mean) frequency F_i as [1]

$$\mathrm{VTL} \approx \frac{(2i-1)c}{4F_i} \tag{1}$$

where c is the velocity of the sound.

From equation 1

$$\frac{\mathrm{VTL}_{adults}}{\mathrm{VTL}_{kids}} \approx \frac{F_{i,kids}}{F_{i,adults}} \tag{2}$$

Therefore normalization of the VTL can be achieved by a frequency warping procedure.

According to [6], MFCC coefficients [8] give better recognition performance with children than LPC's, because of the mel scale frequency warping. However the mismatch between male, female and kids remains. A VTL dependent warping is still necessary to obtain VTL-normalized speech parameters. Most papers indicate that the use of frequency warping procedures improves Speaker Independent Speech Recognition performance with Males and Females [4][7] . [7] gives an efficient frequency warping method for the calculation of normalized MFCC's. Here the concept of frequency warping and the use of MFCC's is combined [4] [7].

3 ESTIMATION OF THE HIGH FREQUENCY POWER SPECTRUM

When the power spectrum of the input speech is scaled downwards to compensate for the VTL differences between adults and kids a gap in the high frequency region is generated. This affects the accurate computation of the mel spectrum and in particular its high frequency part. To avoid this undesirable effect three different strategies have been adopted and have been compared against each other, namely the time domain interpolation, the extrapolation and the oversampling.

3.1 TIME DOMAIN INTERPOLATION

The input signal is interpolated in the time domain so that its spectrum is broadened. Here a simple interpolation of factor two has been employed. Thereafter the enlarged spectrum is scaled downwards by an appropriate factor which is of the order of 1.2. Obviously the generated gap in this case ranges between $.8F_s$ and F_s which is outside the useful region (0-$.5F_s$), (F_s stands for the original sampling frequency). Hence, no inaccuracies during the computation of the mel spectrum coefficients should occur, except for those which are due to the interpolation technique itself. The latter is based on simple averaging of neighboring samples.

3.2 LOW PASS FILTERING EXTRAPOLATION

Here a more original approach has been implemented. Since phase is of no use for mel spectra only the power spectrum is considered and extrapolated before any shrinking occurs. For that purpose a very simple technique is used based on a low pass filtering of the power spectrum itself, treated as a regular signal in this case. The spectrum is extrapolated up to the frequency $.6F_s$ so as to guarantee that no high frequency gap occurs after warping by factor 1.2 as it is shown on fig. 2 and fig. 3, where point 256 corresponds to $.5F_s$. The lowpass filter is an 8th order Chebyshev zero phase with cutoff frequency at $.1F_s$. It was found experimentally that the method is rather insensitive with respect to the cutoff frequency of the filter.

3.3 OVERSAMPLING

This approach is used as a reference to compare the results of both previous methods. Here the input signal is oversampled so that after warping the high

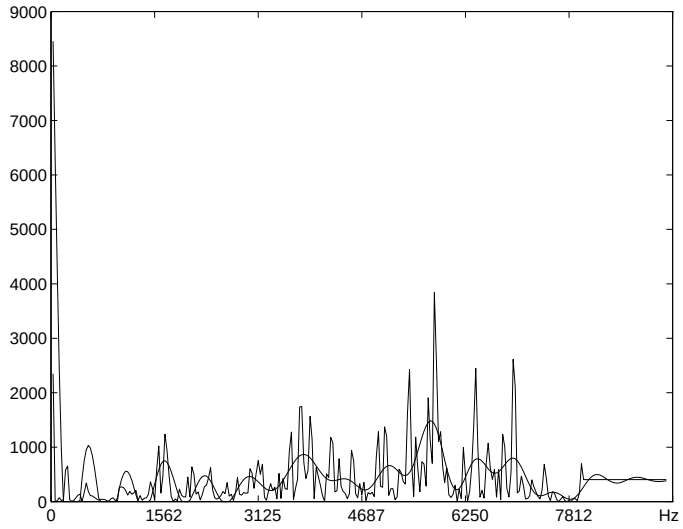

Figure 2: The power spectrum of a speech frame and its lowpass filtering extrapolation, (smoothed curve).

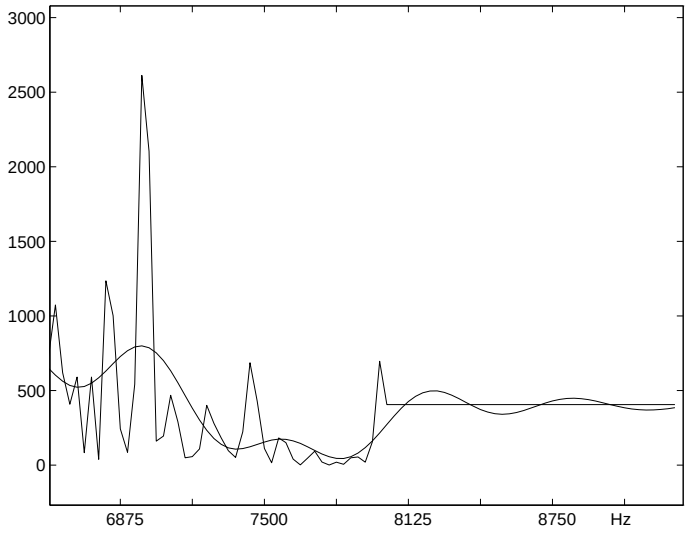

Figure 3: A zoom-in view of figure 2 at high frequencies

frequency gap occurs beyond .5F$_s$. Hence the calculation of the mel spectra which is limited to $.5F_s$ is not influenced by the warping operation. Therefore the results of this method provide an upper bound for the performance of both previous techniques and have been used for comparisons.

3.4 RECOGNITION EXPERIMENTS

All tests were performed on continuous word recognition using Continuous Density HMM's with 1 model per word (11 states). In fact for this particular task when the models are trained on kids and tested on kids the error rate is only 1.0% as shown in table 1. Interesting enough, when the training data of the adults (males and females) is used instead to train the HMM's and tested on kids, the error rate increases to 4.3%.

models used	Word Error Rate
kidsmodels $(\mu_{kids}, \Sigma_{kids})$	1.0 %
adultmodels (μ, Σ)	4.3 %

Table 1: Recognition results when the system is both trained and tested with kids and when it is trained with adults and tested with kids

Method	Word Error Rate
Interpolation	2.2 %
Extrapolation	1.7 %
Oversampling	1.6 %

Table 2: Recognition results with three different techniques for high frequency information recovery

Table 2 shows the recognition error when the models are trained with adults and tested with kids. The kids input speech is transformed prior to processing by one of the three previously mentioned techniques. Note that oversampling gives the best result, 1.6% recognition error, followed by extrapolation with 1.7%. Interpolation is considerably worse achieving only 2.2%. These results correspond to a warping factor (scaling) of 1.16. These results suggest that oversampling may be avoided and simple extrapolation of the power spectrum can be used without any significant deterioration of the performance. It is also interesting to compare the above with results reported in [11] where the models instead of the input speech are transformed and the overall performance is lower.

4 CONCLUSION

The on-line transformation of the power spectrum of speech presented in this paper adapts the VTL characteristics of a speaker group, to the specifications of an already trained speech recognition system. The new extrapolation scheme which is proposed provides similar results to those achieved by oversampling. An improvement of 50% is observed when the adult trained system is used with children's voice.

References

[1] H. Wakita. Normalization of vowels by vocal tract length and its application to vowel identification. In *IEEE ASSP*, Vol. 25, pp.183, 1977.

[2] G.E.Peterson and H.L.Barney. Acoustic Theory of Speech Production. *The Hague, The Netherlands: Mouton, 1960*

[3] G.E.Peterson and H.L.Barney. Control methods Used in a Study of the Vowels In *the Journal of the Acoustical Society of America*, Vol. 24, No. 2 pp.175–184, 1952.

[4] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *Int. Conf. Acoust. Speech & Signal Processing*, pages 346–348, 1996.

[5] H. Wakita. Estimation of Vocal Tract Shapes from Acoustical Analysis of the Speech Wave: the State of the Art In *IEEE ASSP*, Vol. 27, pp.281, 1979.

[6] J.G. Wilpon and C.N. Jacobsen. A study of speech recognition for children and the elderly. In *Int. Conf. Acoust. Speech & Signal Processing*, 1996.

[7] L. Lee and R.C. Rose. Speaker normalization using efficient frequency warping procedures. In *Int. Conf. Acoust. Speech & Signal Processing*, pages 353–356, 1996.

[8] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. ASSP–28, No. 4 pp.357–366, 1980.

[9] G. Fant. Speech Sounds and Features. Cambridge MA: The MIT Press., 1973.

[10] M.J.F. Gales and S.J. Young. Cepstral parameter compensation for HMM recognition in noise. In *Speech Communication*, Vol. 12, No. 3 pp.231–240, 1993.

[11] T. Claes I. Dologlou L. ten Bosch and D. Van Compernolle. New transformations of cepstral parametres for automatic VTL normalization in speech recognition. In *Eurospeech 97*, Vol. 3, pp.1363-1366.