HYBRID RECOGNIZER OF ISOLATED WORDS

Karel Veselý
Bachelor Degree Programme (1), FIT BUT
E-mail: xvesel39@stud.fit.vutbr.cz

Supervised by: František Grézl
E-mail: grezl@fit.vutbr.cz

ABSTRACT

This paper deals with recognition of isolated words by hybrid approach. The brief description of the principles of hybrid recognition is introduced, the recognition phases and feature optimisations are introduced. The report is concluded by practical part with performance results.

1. INTRODUCTION

Automatic speech recognition is a process of converting speech signal to a sequence of words. Well working isolated words recognizer has various practical applications. It can be used to build interactive applications with voice controled user interface. It can also be embeded with benefit in electronical dictionaries, when the word pronounciation is known, but the transcription is not.

The recognition process using hybrid approach has three main phases: feature extraction, estimation of phoneme probabilities and decoding. The core of hybrid approach is utilisation of Artificial Neural Network (ANN) as estimator of phoneme probabilities. These are decoded with standard Viterbi algorithm [1].

Figure 1: Scheme of hybrid recognition

2. FEATURE EXTRACTION

The speech signal is segmented into 25ms long frames with 10ms shift. The spectrum is computed by FFT and integrated by a bank of triangular filters. The triangles are equally distanced in mel-scale, to match acoustical perception of humans. Each triangle represents one „critical band“. We used 15 critical bands for 8KHz and 23 bands for 16KHz data. Output of this filter bank is used as input of phoneme probability estimator.

The speech theory says, that the length of vocal tract directly affects the basal tone from which all the other voice „sounds“ are derieved. The VTLN (vocal tract length normalisation) is a method which compensates this problem. It scales the frequencies of signal wi-

th a single coeficient (typically from interval 0.8 to 1.2). If we find the correct coeficient for each speaker, we can normalize the signal to hit the same frequency interval. But it is not possible to find VTLN factor analytically. The log likelihood of model for several factor has to be evaluated and then the one giving the best likelihood is taken. The *Force alignment algorithm* [1] is used for logarithmical likelihood computation.

For better statistical characteristics of speech parametres, mean and variance normalisation of each coeficient of parameter vector is used. The normalisation is done for each speaker separately.

3. PHONEMES ESTIMATION

The phoneme probability estimator is the heart of hybrid speech recognizer. It transforms parameters to the probabilities of phonemes. We are using set of 45 phonemes, each with 3 states, thus the estimator has 135 outputs.

The phoneme probability estimator can be more complex than just one ANN. In our case a hierarchical structure of three ANNs is used. Two are contextual NN which see context of 15 frames to the past or future together with the actual frame, the third one is merging the partial outputs to form final probabilities. For further details about estimators see [2], block diagram is shown on figure 2a).

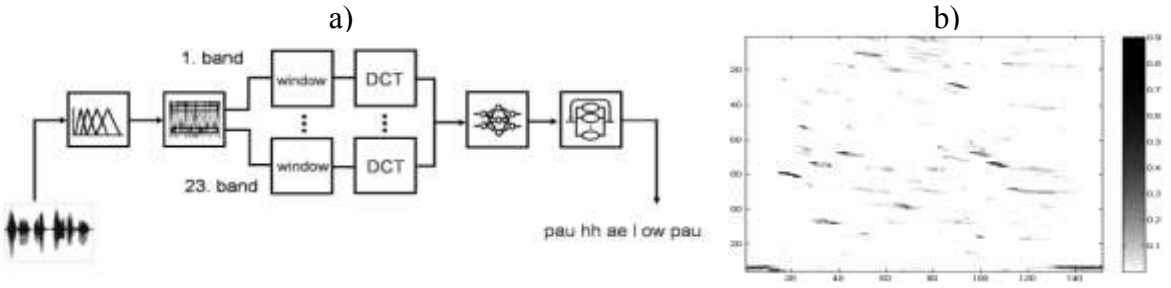

Figure 2: a) Hybrid recognizer sheme, b)Estimated phoneme probabilities (word 'meteorological')

4. RESULTS

Speech database of 10 english speakers was used for experimets. 995 words has been recorded from each speaker. The spoken words are the same for all speakers. This database was provided by the Lingea company.

So far I have made 3 experiments with different setups. The speech was downsampled at 8KHz for the first experiment. Recognition was performed with mean-variance feature normalisation. Used phoneme probability estimator was trained on 300h of 8KHz telephone speech data. For results see: Table 1, row 1

The second setup was for 16KHz sampling rate, with estimator trained on 30h of 16KHz „meeting speech“ data. There were not used any feature optimisations. Results are in Table 1, row 2.

The third experiment was performed for 16KHz sampling rate, with estimator trained on the same data as in previous case. The difference is that in the training procedure was used

VTLN feature optimisation, so that it should be also used in recognition. The VTLN factors were estimated from correctly recognized words only.

	Word error rate (WER) [%]											
Speaker	**I.a**	**I.b**	**II.**	**III.**	**IV.**	**V.**	**VI.**	**VII.**	**VIII.**	**IX.**	**X.**	**AVG**
8KHz – normalization	96,9	91,6	95,2	96,8	91,3	98,5	90,7	88,7	74,8	89,0	84,3	90,7
16KHz – No optimisation	42,3	40,6	35,2	41,0	42,9	78,8	26,5	38,0	39,0	33,0	59,1	43,3
16KHz – VTLN	43,2	34,7	25,9	40,7	42,5	78,4	25,0	34,0	41,1	32,1	64,8	42,0

Table 1: Recognizer performance

5. CONCLUSION

Altough the performance in 8KHz was poor, it is still better than random probability of the correct recognition, which is 0,1% WER (word error rate). The probable reason is data mismatch between training and test data. The telephone band is 300 – 3400Hz, whereas our data are in full band 0 – 4000Hz. The results with 16KHz recognizers are much better than 8KHz. The usage of VTLN is reflected in reduction of WER by 1,3%.

Though the performance of the system is not great, it is possible to use it in a dictionary application, because it is supposed to work in N-best mode. This means that the recognizer will output N most probable word hypotheses and the user can choose the one he wants. It is much more likely to get the correct word this N-tuple.

The 16KHz recognizer with VTLN and mean-variance feature normalisation will be built next. The way of estimating VTLN factor will also change, so far it has been estimated from the log likelihoods of the correctly recognized words, now it will be computed from all the words via *force alignment algorithm* [1]. Focus will be put on dependency of VTLN factor estimation on amount of speech data.

ACKNOWLEDGMENTS

This work was partly supported by Grant Agency of Czech Republic under project No. 102/05/0278 and by Czech Ministry of Education under project No. MSM0021630528. The hardware used in this work was partially provided by CESNET under projects No. 119/2004, No. 162/2005 and No. 201/2006.

REFERENCES

[1] S. Young, J. Jansen, J. Odell, D. Ollason, and P.Woodland.: The HTK book. Entropics Cambridge Research Lab., Cambridge, UK, 2002.

[2] P. Schwarz, P. Matějka, J. Černocký: Hierarchical structures of neural networks for phoneme recognition, In: Proceedings of ICASSP 2006, Toulouse, FR, 2006

[3] P. Schwarz, P. Matějka, J. Černocký: Towards Lower Error Rates in Phoneme Recognition, In: Proceedings of 7th International Conference Text, Speech and Dialoque 2004, Brno, CZ, Springer, 2004