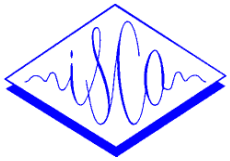

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

HIGH PERFORMANCE CONNECTED DIGIT RECOGNITION THROUGH GENDER-DEPENDENT ACOUSTIC MODELLING AND VOCAL TRACT LENGTH NORMALISATION

Ramalingam Hariharan and *Olli Viikki*

Nokia Research Center, Speech and Audio Systems Laboratory, Tampere, Finland
Email: {ramalingam.hariharan, olli.viikki}@nokia.com

ABSTRACT

Large inter-speaker variability of speech is one of the major sources which degrade the performance of state-of-the-art speech recognition systems. During the recent years, several methods, including gender-dependent acoustic modelling and vocal tract length normalisation, have been developed to reduce this variability. In this paper, we first investigate these two methods individually and propose how they should be implemented in real-world speech recognition systems. Secondly, we show that by combining these two techniques, it is possible to further reduce the error rate in a connected digit recognition task under a realistic car noise environment. Experimental results justify the use of the combined approach. A 44.1% decrease in string error rate was observed when the performance of the joint system was compared to the gender-independent baseline system. The results were also better than that obtained when using these techniques individually.

1. INTRODUCTION

Large inter-speaker variability is one of the most fundamental problems in speaker-independent speech recognition which restricts both the performance and utilisation of practical Automatic Speech Recognition (ASR) systems. While an acceptable recognition accuracy can usually be obtained with the majority of speakers, there are always a significant proportion of so called outlier speakers for whom the recognition rate obtained does not meet the performance objectives.

Inter-speaker variability includes physiological differences across various speakers, e.g. the variation of the vocal tract length and shape, as well as various speaking habits and pronunciation differences, e.g. different dialects and the linguistic variations of non-native speakers. Since the state-of-the-art speaker-independent speech recognition technology relies heavily on the power of the statistics, small pronunciation deviations cannot appropriately be considered in the acoustic models, e.g. continuous density HMMs, whose parameters have been averaged over speakers in the training set. This inappropriate modelling results in recognition performance degradation.

Various kinds of HMM adaptation techniques (e.g. [10]) have been developed over the years to reduce inter-speaker variability [3][10]. While these methods try to alter the parameters of the acoustic models, speaker normalisation methods modify only the observations i.e. feature vectors that are presented to the recognition unit. The simplest form of speaker normalisation is cepstral mean normalisation, where the long-term cepstral mean estimate is subtracted from computed cepstral parameters. An alternative speaker normalisation technique called Vocal Tract Length Normalisation (VTLN) [1][8] has recently produced promising results. VTLN techniques aim at modifying the frequency scale so that vocal tract length differences between various speakers are compensated.

Knowing that the recognition error rates for males and females are different, another straightforward method for reducing the inter-speaker variability is to utilise gender specific acoustic models. Gender-dependent HMMs have shown to help in reducing the speaker variability [14]. Gender-specific acoustic modelling can be regarded as a speaker clustering technique where speakers with similar speaking characteristics are pooled together, and separate models sets are generated for each speaker cluster. A definite advantage of gender-dependent modelling to other speaker clustering techniques is the fact that the speaker allocation into two clusters is pre-defined. In this section, we propose to combine VTLN and gender-dependent modelling techniques to be combined for maximising the acoustic modelling accuracy.

The remainder of this paper is organised as follows. The basic outline of the approaches used for vocal tract length normalisation and gender dependent modelling in this paper are given in the next two sections. The details of the experiments are given in the next section. The final section presents a summary of the paper.

2. GENDER DEPENDENT MODELLING

It is widely known that there is a great deal of variation between male and female speech. This can be observed for example in recognition tests as the error rates for males and females are typically very different from each other, see e.g. [6]. A straightforward and widely used idea is to build up a separate set of acoustic models for both genders. Gender-dependent acoustic modelling can also be regarded as a speaker clustering technique, even though the speaker allocation is predefined. Due to many problems related to more analytical and complex clustering techniques, e.g. the availability of the training data for several clusters, gender-dependent modelling is often a very viable option for speaker clustering and model training.

There are two alternatives how gender-dependent modelling can be realised in practice. If the speaker's sex is known, or it can somehow be detected, one can easily select an appropriate set of acoustic models to be active during recognition. Alternatively, it is possible to keep both male and female model sets active and let the recogniser decide which models are used in decoding. As shown in Section 4, this automatic approach can produce a lower error rate. This is due to the fact that the distinction between male and female voices is not as straightforward as gender detection, i.e. female speech is not always best characterised by female models. By letting the recogniser to

select the models among male and female models, one is capable of choosing the HMM according to the Maximum Likelihood (ML) criterion. An obvious disadvantage of the recogniser-driven selection approach is related to the increased implementation costs. Both memory consumption and computational complexity will be doubled in this case. This is likely to limit the use of gender-dependent modelling techniques in applications which have sparse memory and processing power resources.

Even though gender-dependent acoustic modelling is easy to implement, provided there is sufficient amount of free resources available, and it reduces error rate, there is still plenty of room for performance improvements. Therefore, gender-dependent modelling needs to be combined with some other compensation techniques for achieving the best possible recognition performance. Viable methods to be combined with gender-dependent acoustic modelling are for example HMM adaptation techniques or vocal tract length normalisation as shown in this paper.

3. VOCAL TRACT LENGTH NORMALISATION

As mentioned previously, VTLN attempts to normalise the speech signal to an average vocal tract length, so that the parameterised speech signal is independent of inter-speaker differences. The basic VTLN algorithm used in this paper is based on [8,10]. Vocal tract length normalisation is realised by introducing an additional warping factor α to the mel-scale definition as follows

$$f_{\text{mel}} = 2595 \cdot \log_{10}\left(1 + \alpha \frac{f}{700}\right) \quad (1)$$

The normalisation task is then to find an optimal value of α for each speaker and input utterance over a set of pre-defined α values. In [10], it was proposed to select the warping factor which provided the highest likelihood for the given input utterance

$$\hat{\alpha} = \arg\max_{\alpha} P\left(X(\alpha) \middle| \alpha, \theta, W\right) \quad (2)$$

The possible values of α varied between 0.88 and 1.12 corresponding to a 25% variation in the vocal tract length or formant location.

VTLN can be applied either during both the training and recognition phases, or during the training or recognition phases only. The disadvantage of applying VTLN during the recognition phase is an increase in the computational complexity associated with the search for an optimal warping factor for each speaker and/or utterance. However, the reduction of inter-speaker variability during the training phase alone does not necessarily guarantee any increase in the recognition performance as observed previously [e.g. 5]. The advantage of applying speaker normalisation techniques, such as VTLN or speaker adaptive training [1], during training phase is a faster adaptation speed in speaker adaptation as compared to the use of uncompensated models. It is thus necessary to apply compensation during the recognition phase to increase the recognition performance of a target speaker. However, this leads to an increase in the computational complexity.

A high run-time computational complexity is a clear problem in VTLN as one needs to process the input utterance multiple times with the different values of α. This restricts the utilisation of VTLN in real-world speech recognition. We recently proposed a speaker-specific VTLN technique [5] where a single warping factor value was searched for a target speaker. This arrangement enabled a huge reduction in the run-time complexity since the warping factor for each speaker was computed off-line from a small number of utterances spoken by a target speaker. These utterances were noise-free in order to avoid a warping factor estimate, which is affected by the effects of background noise and other disturbances. After the optimal warping factor has been found based on these front-end "adaptation" utterances, the warping value was then fixed and used for recognising all the utterances spoken by the speaker.

This speaker-specific VTLN was not found to degrade the performance when compared to utterance-specific warping factor search. Our assumption was that one does not necessarily need to search a separate warping factor for each input utterance since the vocal tract shape and length are static speaker-specific characteristics that vary very little utterance by utterance. No degradation in performance is observed if the speaker-specific warping factor is robust and it has been estimated over many enough utterances.

In this paper, we realised the speaker-specific VTLN approach as a system controlled enrolment session during which the system learns the speaker's pronunciation characteristics and modifies the feature vector stage accordingly.

4. EXPERIMENTAL EVALUATION

The experimental evaluation of the above methods was carried out for a speaker-independent connected digit recognition experiment based on whole-word HMMs.

4.1 Experimental Details

A separate Finnish language connected digit database was used for HMM training. The following four sets of speaker-independent digit HMMs were created from the training database:

- Gender-independent models
- Gender-independent, vocal tract length normalised models
- Gender-dependent models
- Gender-dependent, vocal tract length normalised models

The training set consisted of utterances spoken by 375 speakers in a variable car noise environment. For multi-environment type of speaker-independent training, all utterances were pooled together and a set of state duration constrained digit HMMs [8] (2 Gaussian mixtures per state) were estimated according to the ML criterion. The number of male and female speakers, the number of digits, and the transitions between the digits were balanced in the training database.

The testing database used in digit recognition evaluations was a Finnish connected digit database recorded in a car environment. This database consisted of 9 speakers (4 males and 5 females) with each speaker speaking at least 1,000 utterances. Recordings

were carried out during four recording days over a time-span of one month. Each recording session took approximately one hour during which a test speaker spoke connected digit triples in continuously changing noise conditions depending on the speed of the car, road, and traffic conditions. In the off-line recognition experiments, the order of the test utterances was further randomised so that consecutive utterances were *not* necessarily spoken in similar noise conditions. This arrangement enabled us to simulate a real-life usage pattern. In addition to these test utterances, each speaker also uttered 30 connected digit triples in a noise-free car environment. These utterances were used in the VTLN experiments to find the best speaker-specific warping factor for each test speaker.

A feature vector set consisting of 39 coefficients, 12 Mel Frequency Cepstral Coefficients (MFCC), log-energy, and their first and second-order time derivatives, was used for all the experiments in this paper. These feature vectors were further normalised to have similar parameter statistics in all noise conditions as described in [13]. Table 1 summarises the results achieved with this baseline system in connected digit recognition. The baseline system did not include any compensation schemes.

Digit String Recognition Accuracy [%]		
Male	**Female**	**Average**
88.41	71.06	77.51

Table 1. Recognition rates obtained with the baseline system.

4.2 Gender-Dependent Experiments

Experiments were next carried out to determine the best approach to realise gender-dependent model selection for maximising the recognition performance. The previously described two approaches were experimentally compared. The gender detection was assumed to be optimal in these tests, so that the correct model set was always applied to the correct sex. Table 2 shows the results of the experiments.

Method	**Digit String Recognition Accuracy [%]**		
	Male	**Female**	**Average**
Baseline	88.41	71.06	77.51
Gender selection	82.30	71.24	75.33
Recogniser selection	91.68	75.47	80.99

Table 2. A comparison of two different gender-dependent acoustic modelling approaches.

It is seen from Table 2 that the use of separate gender specific models for different genders performed worse than even the baseline results. As explained earlier, this could be due to the fact that the gender specific models may not always characterise the speech of the particular gender properly. The use of both the gender-specific models together for recognition produced the best performance, with a 15.5% decrease in the string error rate over the gender-independent baseline results.

4.3 Gender-Independent VTLN Experiments

The results of applying VTLN to gender-independent (GI) models are given in Table 3. The result of applying VTLN during testing separately is given in the second row. The final row presents the results when VTLN was applied during both training and testing.

As shown in Table 3, there is an improvement in the recognition accuracy (error rate decrease of more than 7%) over the baseline when VTLN was applied during testing alone. If speaker normalisation was applied during both training and testing (utterance-wise warping), over 21% string error rate decrease was obtained. The recognition rates are also presented separately for males and females. The tests produced relatively higher increase in recognition accuracy for females than for males, when VTLN is used both for training and testing.

VTLN Train	**VTLN Test**	**Digit String Recognition Accuracy (%)**			**Error Rate Reduction (ERR %)**
		Male	**Female**	**Average**	
No	**No**	88.41	71.06	77.51	-
No	**Yes**	90.18	72.56	79.14	7.25
Yes	**Yes**	89.68	77.91	82.29	21.25

Table 3. Recognition results on applying speaker-specific VTLN with gender-independent HMMs.

4.3 Gender-Dependent VTLN Experiments

The results of the VTLN experiments using the gender-dependent (GD) models are summarised in Table 4. Gender-dependent acoustic modelling appeared to be an efficient method to cope with speaker variability because it resulted in a 15.5% digit string error rate reduction compared with the baseline system.

The results in Table 4 also show that the use of VTLN on top of gender-dependent HMMs further improved the recognition accuracy. The addition of speaker-specific VTLN during the training and recognition phases produced a significant improvement of approximately 34% decrease in the digit string error rate compared to the case when VTLN was performed on the gender-dependent models. This could mean that the effect of using gender-dependent models and vocal tract length normalisation are essentially additive. As shown in the table, there was a 44.1% decrease in string error rate with the combined approach compared to the use of gender-independent models. It was also noticed that the use of VTLN during training and the recognition phase produced higher recognition accuracy than its use only during testing, similar to the results obtained with gender-independent models.

VTLN Train	**VTLN Test**	**Digit String Recognition Accuracy (%)**			**ERR(%) w.r.t GI Baseline**	**ERR(%) w.r.t GD Baseline**
		Male	**Female**	**Aver.**		
No	**No**	91.68	75.47	80.99	15.47	-
No	**Yes**	95.10	76.37	83.33	25.88	12.31
Yes	**Yes**	96.55	82.04	87.43	44.11	33.88

Table 4. Recognition results on applying speaker-specific VTLN with gender-dependent HMMs.

4. CONCLUSIONS

This paper examined the use of two different speaker variability reduction techniques – gender dependent modelling and vocal

tract length normalisation, for maximising the recognition performance in a connected digit recognition task under a car noise environment. The use of gender-specific models alone produced a 15.5% decrease in the string error rate for a Finnish connected digit recognition task when compared to the use of gender independent models. We also showed that the combined use of both sets of gender-specific models produced better performance than the use of separate sets of gender-specific models depending on the gender of each speaker. The combination of previously proposed speaker-specific Vocal Tract Length Normalisation (VTLN) and gender-dependent modelling produced a further improvement in performance with a 44.1% decrease in the string error rate, indicating that the effect of using VTLN and gender dependent modelling could be essentially additive. These results were also better than that obtained for speaker specific VTLN with the use of gender-independent models.

REFERENCES

1. Anastasakos, T., McDonough, J., Schwartz, R., Makhoul, J., 1996. "A compact model for speaker adaptive training", Proceedings of International Conference on Spoken Language Processing, pp. 1137-1140.
2. Eide, E., Gish, H., 1996. "A parametric approach to vocal tract length normalisation", Proceedings of International Conference on Acoustics Speech, and Signal Processing, pp. 346-348.
3. Gauvain, J., Lee, C.-H., 1994. "Maximum a Posteriori estimation of multivariate Gaussian mixture observations of Markov chains", IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 2, pp. 291-298.
4. Gong, Y., Godfrey, J.J., 1999. "Transforming HMMs for speaker-independent hands-free speech recognition in the car", Proceedings of International Conference on Acoustics Speech, and Signal Processing
5. Hariharan, R., Viikki, O., 1999., "On Combining Vocal Tract Length Normalisation and Speaker Adaptation for Noise Robust Speech Recognition", Proceedings of Eurospeech 99, Vol. 1, pp. 215-218, Budapest, Hungary.
6. Hariharan, R., Viikki, O., 2000. "An integrated study of speaker normalisation and HMM adaptation for noise robust speaker-independent speech recognition", Submitted to Speech Communication journal.
7. Hunt, M.J., 1999. "Some experience in In-Car Speech Recognition", Proceedings of the Nokia-Cost-IEEE Workshop on Robust Methods for Speech Recognition in Adverse Conditions, pp. 25-31.
8. Laurila, K., 1997. "Noise robust speech recognition with state duration constraints", Proceedings of International Conference on Acoustics Speech, and Signal Processing, pp. 871-874.
9. Lee, L., Rose, R.C., 1996. "Speaker normalisation using efficient frequency warping procedures", Proceedings of International Conference on Acoustics, Speech, and Signal Processing, pp. 353-356.
10. Lee, L., Rose, R.C., 1998. "A frequency warping approach to speaker normalization", IEEE Trans. on Speech and Audio Processing, Vol. 6, No. 1, pp. 451-454.
11. Leggetter, C.J., Woodland, P.C., 1995. "Maximum likelihood linear regression for speaker adaptation of continuous density Hidden Markov Models", Computer Speech & Language, Vol. 9, pp. 171-185.
12. Viikki, O., Laurila, K., 1998. "Incremental on-lin speaker adaptation in adverse conditions", Proceedings of International Conference on Spoken Language Processing, pp. 1779-1782.
13. Viikki, O., Bye, D., Laurila, K., 1998. "A recursive feature vector normalization approach for robust speech recognition in noise", Proceedings of International Conference on Acoustics, Speech, and Signal Processing, pp. 733-736.
14. Woodland, P.C., Leggetter, C.J., Odell, J.J., Valtchev, V., Young, S.J., 1995. "The development of the 1994 HTK large vocabulary speech recognition system", Proceedings of ARPA Spoken Language Technology Workshop, pp. 104-109.