Proc. of Intl. Conf. Acoustic, Speech and Signal Proc., Munich, Germany, April 1997.

ON COMBINING FREQUENCY WARPING AND SPECTRAL SHAPING IN HMM BASED SPEECH RECOGNITION

Alexandros Potamianos and Richard C. Rose

AT&T Labs, Murray Hill, NJ 07974, U.S.A.

ABSTRACT

Frequency warping approaches to speaker normalization have been proposed and evaluated on various speech recognition tasks [1, 2, 3]. These techniques have been found to significantly improve performance even for speaker independent recognition from short utterances over the telephone network. In maximum likelihood (ML) based model adaptation a linear transformation is estimated and applied to the model parameters in order to increase the likelihood of the input utterance. The purpose of this paper is to demonstrate that significant advantage can be gained by performing frequency warping and ML speaker adaptation in a unified framework. A procedure is described which compensates utterances by simultaneously scaling the frequency axis and reshaping the spectral energy contour. This procedure is shown to reduce the error rate in a telephone based connected digit recognition task by 30-40%.

1. INTRODUCTION

A major hurdle in building successful automatic speech recognition applications is non–uniformity in performance across a variety of conditions. Many successful compensation and normalization algorithms have been proposed in the literature dealing with different sources of variability. Typical examples in telecommunications applications of speech recognition include inter–speaker, channel, environmental, and transducer variability. In practice, a speech utterance may be simultaneously affected by many sources of variability, and there may be many acoustic correlates associated with a given source of variability. As a result, it is important that different procedures for compensating for acoustic distortions be tightly coupled with one another. This paper attempts to address how linear model transformation and speaker normalization by frequency warping can be implemented as a single procedure to compensate for these sources of variability.

Model adaptation techniques have been used for improving the match between a set of adaptation utterances and the hidden Markov model (HMM) used during recognition. The parameters of a linear transformation are estimated using a maximum likelihood criterion and the transformation is applied to the HMM parameters [4]. A common problem among these techniques is the existence of speakers in a population whose speech recognition performance does not improve after adaptation. This can be especially true for unsupervised, single utterance based adaptation scenarios. It is generally thought that only those distributions in the model that are likely to have generated the adaptation observations have a chance to be mapped to the target speaker. Therefore, if the "match" between the model and the adaptation utterance is not reasonably "good" to begin with and the number of adaptation utterances is limited, then the utterance cannot "pull" the model to better match the target speaker.

Speaker normalization by frequency warping has been used for estimating a frequency warping function that is applied to the input utterance so that the warped utterance is better matched to the given HMM model. As is the case for model adaptation, there exists a subset of utterances for which frequency warping does not improve performance. The ineffectiveness of speaker normalization for these utterances is thought to be due to the interaction of other sources of variability in the process of estimating the "best" warping function. If both the model adaptation and speaker normalization procedures are limited by the initial relationship between the HMM model and the input utterance, then perhaps a solution to this problem is to search for an optimum warping function and an optimum model transformation in the same procedure. This is the principle focus of this paper.

The paper is organized as follows. First, the frequency warping based speaker normalization procedure is described in Section 2. In Section 3, a combined procedure for frequency warping and model adaptation is described and applied to a single utterance based adaptation paradigm. A discussion of the application of frequency warping as applied to childrens' speech recognition using HMM models trained from adult speakers is given in Section 4. Finally, discussion and summary is provided in Sections 5 and 6.

2. SPEAKER NORMALIZATION USING FREQUENCY WARPING

In [3], an efficient frequency warping algorithm for speaker normalization was proposed and applied to telephone based speech recognition. The frequency warping approach to speaker normalization compensates mainly for inter-speaker vocal tract length variability by linear warping of the frequency axis by a factor α. By applying frequency warping during both training and recognition it was shown that word error rate can be reduced by approximately 20%. The frequency warping algorithm described in [3] is briefly presented next.

Frequency warping is implemented in the mel-frequency filterbank front-end by linear scaling of the spacing and bandwidth of the filters. Scaling the front-end filterbank is equivalent to resampling the spectral envelope using a compressed or expanded frequency range. The speaker normalization algorithm works as follows. For each utterance, the optimal warping factor $\hat{\alpha}$ is selected from a discrete ensemble of possible values so that the likelihood of the warped utterance is maximized with respect to a given HMM and a given transcription. The values of the warping factors in the ensemble typically vary over a range corresponding to frequency compression or expansion of approximately ten percent. The size of the ensemble is typically ten to fifteen discrete values. Let $X^\alpha = g_\alpha(X)$ denote the sequence of