UTTERANCE DEPENDENT PARAMETRIC WARPING FOR A TALKER-INDEPENDENT HMM-BASED RECOGNIZER

Daniel J. Mashao and John E. Adcock[*]

LEMS, Division of Engineering, Brown University, Providence, RI 02912, USA
e-mail: {djm,jea}@lems.brown.edu
ftp://ftp.lems.brown.edu/pub/speech/papers/icsp97.ps.gz

ABSTRACT

In an effort to improve recognition performance of talker-independent speech systems, many adaptive methods have been proposed. The methods generally seek to exploit the higher recognition performance rate of talker-dependent systems and extend it to talker-independent systems. This is achieved by some form of placing talkers into several categories, usually using gender or vocal-tract size. In this paper we investigate a similar idea, but categorize each utterance independently. An utterance is processed using several spectral compressions, and the compression with the maximum likelihood is then used to train a better model. For testing, the spectral compression with the maximum likelihood is used to decode the utterance. While the spectral compressions divided the utterances well, this did not translate into significant improvement in performance, and the computational cost increase was significant.

1. INTRODUCTION

Research in improving the performance of speech recognition systems is ongoing. Despite the successful application of Hidden Markov Models (HMM) for the back-end of speech recognition systems, the search for an optimal feature-set is still a fundamental problem. The feature-set is important since it is supposed to enhance the semantic discriminating information of the speech signal that will be available to the back-end. If the feature-set is more discriminating, the expected recognition performance should be higher.

Various methods have been used in the attempt to improve the performance of talker-independent systems. They are generally called adaptive systems. Adaptive systems often use talker-dependent methods to improve performance of talker-independent systems. Several approaches have been investigated but they usually categorize the talker into one of several groups [1, 2, 3], with the simplest grouping being to divide talkers by gender into male and female. Formant shape, pitch location, or some other characteristic of the speech signal is used to categorize the talker and select a model for recognition.

In [4] we proposed a parameterized warped feature-set for an HMM-based speech recognition system. The warped feature-set outperformed an LPC mel-cepstrum feature-set. In particular, a certain combination of parameters offered the best performance for the whole talker-independent system. In this study, we attempt to see if the parameterized warping may be used in an utterance-dependent fashion. Secondly, we want to use features computed from different warpings to make a single HMM-model. We evaluate whether the performance of this model will be different from that developed from a single spectral warping, as is usually the case. Thus the system would use some form of spectral adaptation to utterances in both the training and test set. This is different from many talker normalization methods in which normalizations are based on the talker rather than the utterances. Hence, techniques such as pitch or formant analysis were not used in the proposed method.

The feature-set is computed as before, and the likelihood from the HMM model is used to estimate the suitability of the spectral warping to the utterance. The idea of using likelihood for vocal-tract normalization is not new, see, e.g. [1]. To evaluate our approach, we first experiment with a discrete HMM-based system, as it speeds up our experiment. However, due to the vector-quantization process, the discrete system is not ideal. The resource-costly semi-continuous system is really preferable, since it models the observations directly, but to expedite the experiment, it was not used here.

2. THE PARAMETERIZED FEATURE-SET

The (α,β) feature parameterization was first discussed in [4, 5] in which a model is specified by two values (α,β). The advantages of the parameterization is that the effect of spectral compression can be calculated. Unfortunately the calculations are compute-intensive, requiring about 120 hours of CPU time per (α,β) for the discrete system. Using the (α,β) parameterization, it was shown that the spectral compression that approximated mel-scale yielded the best recognition performance. This verified that mel-scale compression is indeed better for talker-independent tasks which is already widely accepted. The α and β parameters are related by the equation

$$A \sum_{i=1}^{\alpha} \beta^{i-1} \quad = \quad N/2, \tag{1}$$

where N is the size of a DFT spectrum and A is a constant greater than 1. The parameters α and β can be chosen independently, subject to Equation 1, to achieve a desired

[*] This work was funded by NSF grants MIP-9120843 and MIP-9509505.

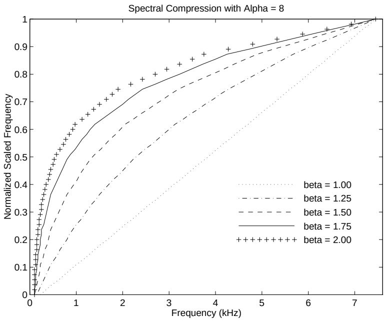

Figure 1. Spectral warping with $\alpha = 8.0$.

spectral compression. In the previously reported experiments [4, 5], α was limited to be between 4 and 10 and β was limited to be between 1 and 2. Mel-scale compression is approximated when $\alpha = 8.0$ and $\beta = 1.5$. Figure 1 shows the spectral compression when $\alpha = 8.0$ and $\beta = 1.0, 1.25, 1.5, 1.75$, and 2.00. It should be noted from Figure 1 that spectral compression is more sensitive to changes in β below 1.5 than above, due to the nonlinearity of Equation 1. In this study an investigation is made to see if the parameterization is in fact utterance-dependent. Although it is generally expected that the dependence would be talker dependent, this condition is not enforced. Thus each utterance is treated independently. The rationale is to find out what kind of model would be generated, and, hopefully, what can be learned from such a model.

3. EXPERIMENTAL TECHNIQUE

3.1. Experimental Set-up

Experiments are performed on a connected, alphadigit, talker-independent HMM-based recognition system. The vocabulary consists of the English alphabet (**A** $\sim$ **Z**), ten digits (**0** $\sim$ **9**), and control words *SPACE* and *PERIOD*. This vocabulary is ideal for testing the discriminating power of the feature-set due to highly confusable words such as in the E-set (**B, C, D, E, G, P, T, V, Z, 3**), the A-set (**A, H, J, K, 8**), the nasal-set (**M, N**), and other word combinations. Most of the differences among the E-set words arise from the initial consonants; for example, the words (**B,D**) are the voiced counterparts of the unvoiced consonants (**P,T**) respectively. On the other hand **B**, **D** and **G** differ based on the place of articulation. These short-time differences make the E-set highly confusable. There are similar consonantal differences in the A-set and the nasal-set.

There is a total of 38 words in the database, and since the recognizer does not use a grammar, it has a perplexity of 38 (actually a little higher due to silences) which is higher than for many large vocabulary systems using bigram or trigram language models. The database consists of speech from 80 male and 40 female native American-English talkers with about 12 hours of speech. It is divided into training (80 talkers), "test and modify" (20 talkers), and testing sets (20 talkers). The features for the HMM-based back-end are classified using three independent codebooks. The codebooks are 1) the direct cepstral coefficients c_n, 2) cepstral difference coefficients, and 3) energy and its difference. The difference coefficients are computed by taking the difference of direct coefficients of two frames that are a frame apart. The HMM uses explicit-duration modeling techniques to improve recognition performance. A K-means vector quantization is used to classify into 256 classes for each of the codebooks.

The best recognition performance obtained on this database has an error rate of 8.2% for the semi-continuous HMM-based system and about 12% for the discrete system. Both error rates were obtained without the use of language models. This high error rates indicates there is still room for improvement in either the feature-set or the back-end.

3.2. Algorithm

The advantage of using Hidden Markov Models in the back-end of the system is the mathematical tractability of the algorithm. In [6] three problems are stated and the first problem is relevant to the algorithm proposed here. The problem is how to efficiently compute $P(O|\lambda)$, the probability of the observation sequence $O = (o_1, o_2, ..., o_T)$ given the model $\lambda = (A, B, \pi)$. The method used to compute the probability is called the *forward-backward algorithm.* In the *forward* path the forward variables $\alpha_t(i)$ (not to be confused with α and β of the feature-set parameterization; the parameterization variables do not have subscripts), is defined as

$$\alpha_t(i) = P(o_1, o_2, ..., o_t, q_t = i|\lambda) \quad (2)$$

that is the probability of the partial observation sequence $(o_1, o_2, ..., o_t)$ up to time t and being in state i given the model λ. The backward variable $\beta_t(i)$ is defined as

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, ..., o_T, q_t = i|\lambda) \quad (3)$$

where T is the number of observations in an utterance. The advantage of these definitions is that the probability of being in state i can easily be computed. Of interest to us here is the observation sequence $O = (o_1, o_2, ..., o_T)$. The observation sequence depends on the front-end signal processing, which also determines the space of the features.

By varying the values of the (α, β) pair the observation sequence is affected, which in turn affects the forward and backward variables. Using the above definition for $\alpha_t(i)$ it can be seen that $\alpha_T(i)$ is the probability that the whole observation sequence is produced by the model λ. A higher value of $\alpha_T(i)$ therefore gives the likelihood that observations O (and in turn the parameterizations) are more suitable to the model λ. Usually for talker-independent systems, without any adaptation, a single best compression may be found for all the talkers in the training set. In [4] we showed that mel-scale compression is a good choice in such applications. On the other hand, for adaptive systems some characteristic of the talker may be used to categorize all the talker's utterances into some class. Both methods have been applied successfully. In this paper, we are proposing a novel method where there are no categories used other than the likelihood of an utterance to select a particular parametric warping. It is, of course, expected that utterances from a single talker will have a similar or nearly similar warpings, but this is not enforced.

The system determines the warping factors based only on the utterance. Figure 2 shows the algorithm for training and testing. In order to make the problem manageable,

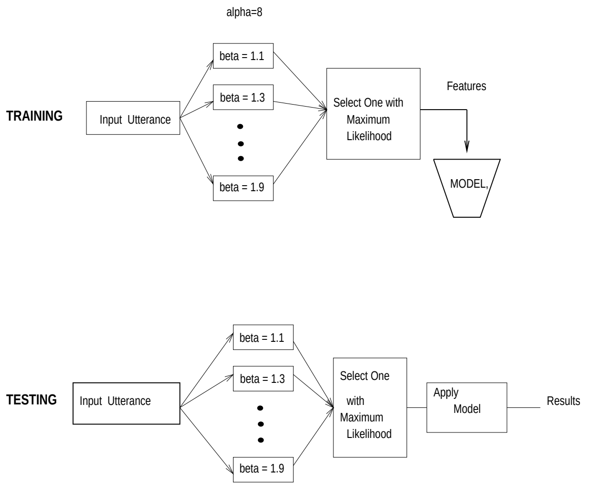

Figure 2. The algorithm for training and testing.

the value of $\alpha = 8$ was fixed, and likelihoods computed for various β values. For training only, the β value producing the highest likelihood was included, and the resulting parameters contributing to the common model. The next iteration model λ is then composed of the best warpings of the previous model. Our initial expectation was that the warpings will cluster near $\beta = 1.5$. We expected to learn from any differences. For testing, each utterance was run through the common model and that with the highest likelihood accepted as the recognized string.

3.3. Computational costs

It is evident that our idea would have been tested better if a continuous HMM-based system had been used, and it was our intention to apply the algorithm on such a system. However, to compute one value in the parametric space required over 120 CPU-hours for the discrete system, while for the semi-continuous system it requires 480 CPU-hours[5]. In addition the algorithm we are proposing requires that a single point be computed for each β value, dramatically increasing the CPU-time by the number of different β's to be computed. We were thus compelled to experiment with only the sub-optimal discrete system. There is further work to see if the best parameterization can be determined before the computations are to be done. If this works out, it will enable experiments on a continuous HMM-based system.

4. EXPERIMENTS

This section describes the results of the experiments performed using the proposed algorithm. Pseudo-random numbers are used to initialize the zeroth model. The first iteration model is trained from the zeroth model. Several utterances from talkers in the training set are used to generate the codebooks. Three codebooks are generated as described earlier. Two methods were used to select the feature-sets to generate codebooks. In the first method, codebooks were developed from observations for spectral warpings of $\beta = (1.1, 1.3, 1.5, 1.7, 1.9)$. This was done to remove any bias towards a particular value of β. When the results of the spectral warpings showed preference to 1.7 and 1.9, we thought the results could have been influenced by

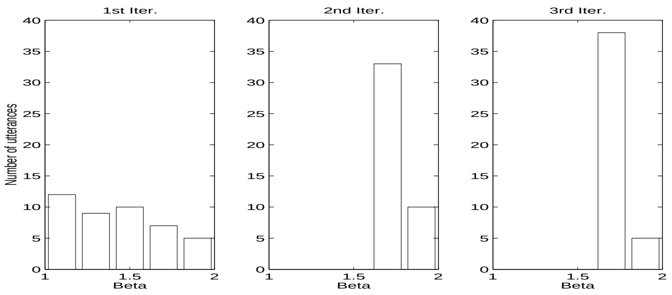

Figure 3. Warping selected for utterances of a female talker through iteration 1 to 3.

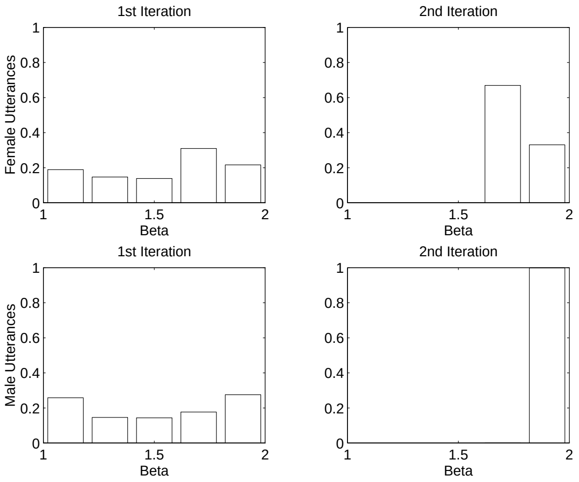

Figure 4. Warping selected for utterances versus gender through first two iterations.

the mixture of the spectral warpings in the codebooks. The experiment was repeated with codebooks that were used to generate previously reported results in [4] where $\beta = 1.5$. The same results were obtained, indicating that the codebooks did not affect the results.

The model training program was modified to read PCM files and generate features using various values of β. This meant that more resources are required to compute the features and vector quantize them on the fly. The warping factor β with the maximum likelihood was recomputed, and its parameters contributed to making the next iteration model. In this way, only the most likely features (warpings) of an utterance are used in building the model. Figure 3 shows which warping factors of $\beta = 1.1, 1.3, 1.5, 1.7,$ and 1.9 were selected in the first three iterations from a pseudo-random model for a female talker. Essentially, by the second iteration 88% of her utterances have selected $\beta = 1.7$ as the best compression.

The results in Figure 3 are typical for females in the database. For most males, by the second iteration almost all the utterances have selected $\beta = 1.9$. The selected β spacing is too coarse to show what is taking place between 1.7 and 1.9. Figure 4 shows the percentage of utterances selecting a β value based on gender during the first two iterations.

After the spread on the compressions was determined,

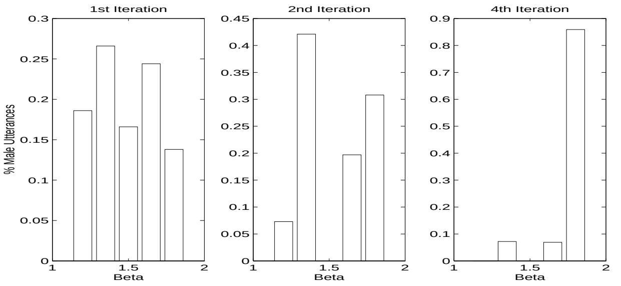

Figure 5. Warping factors selected for utterances using only male talkers.

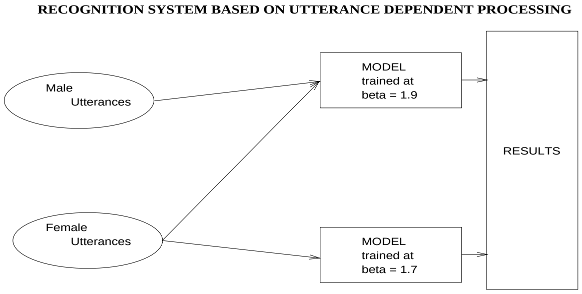
RECOGNITION SYSTEM BASED ON UTTERANCE DEPENDENT PROCESSING
Male Utterances
Female Utterances
MODEL trained at beta = 1.9
MODEL trained at beta = 1.7
RESULTS

Figure 6. Recognition system taking advantage of utterance dependent warping

further experiments were performed on the database. Instead of training a model with both male and female talkers in the utterance-dependent warping algorithm, they were separated. Figure 5 shows the β spread for data trained only on male talkers in the training set. The first iteration of Figure 5 is as expected, since the model is not particularly set to any spectral compression. By the second iteration, an interesting phenomenon happens: almost half of the utterances selects the β value below 1.5 and the other half select above. Also to be noted the value of 1.5 itself is hardly selected. The third iteration is not shown, but by the fourth iteration the spread is similar to the one that was obtained when both the male and female talkers were included in the training set. This shows that somehow when the β is allowed to vary higher values are selected.

After four iterations the new model had a performance of 86.0%. This is slightly better than LPC mel-cepstrum performance of about 85.8% but 2% less than the reported[4] performance of the parameterized front-end. However, mel-cepstrum and parameterized results were from the twenty-fourth iteration model. Thus this performance is expected to increase. But it should also be noted that about the same amount of work is spent on the four iterations in the proposed algorithm as in the twenty-four of [4]. The β spread between male and female talkers can be exploited by using a recognition system as shown in Figure 6. Whether the utterance is evaluated at a specified β will depend on the likelihood. This system is not tested as yet.

4.1. Conclusions

This is work in progress, and more results will be reported at the meeting. The next step is to investigate the warping factor space β between 1.7 and 1.9, and to compute the results on the semi-continuous system. The advantage of a semi-continuous system is that speech features (observations) are used directly to compute the HMM model parameters. This should give a better performance and test of the proposed method. We have achieved our initial goal of investigating whether the spectral compression is utterance dependent. While most female utterances selected $\beta = 1.7$, over 30% selected $\beta = 1.9$, as did most of the male utterances.

The advantage of the proposed system is that no classification of the talker is required, but the utterances themselves are used to obtain an optimal model. The disadvantages, however, are significant. Based on the results of the discrete HMM-based system, it seems the cost of the algorithm is excessive and not worth the small projected performance increase.

REFERENCES

[1] A. Andreou, T. Kamm, and J. Cohen. Experiments in vocal tract normalization. In *Proceedings of the CAIP Workshop: Frontiers in Speech Recognition II*, volume 1, 1994.

[2] Li Lee and Richard C. Rose. Speaker normalization using efficient frequency warping procedures. In *Proceedings of the ICASSP-96, Atlanta,GA*, volume 1, pages 353–356, May 1996.

[3] Steven Wegmann, Don McAllaster, Jeremy Orloff, and Barbara Peskin. Speaker normalization on conversational telephone speech. In *Proceedings of the ICASSP-96, Atlanta,GA*, volume 1, pages 339–341, May 1996.

[4] Daniel J. Mashao. Experiments on a parametric non-linear spectral warping for an HMM-based speech recognizer. *Proceedings of the IEEE ICASSP-96, Atlanta, GA*, May 1996.

[5] Daniel J. Mashao. *Computations and Evaluations of an Optimal Feature-set for an HMM-based Recognizer.* PhD thesis, Brown University, Providence RI, USA, May 1996.

[6] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition.* Englewood Cliffs,NJ:Prentice-Hall, 1993.