A MODEL DISTANCE MEASURE FOR TALKER CLUSTERING AND IDENTIFICATION

J. T. Foote[1] *H. F. Silverman*[2]

[1]Cambridge University Engineering Department, Cambridge CB2 1PZ, UK
[2]Division of Engineering, Brown University, Providence, RI 02912, USA

ABSTRACT

This paper describes methods of talker clustering and identification based on a "distance" metric between discrete HMM output probabilities. Output probabilities are derived on a tree-based MMI partition of the feature space, rather than the usual vector quantization. The information divergence between speaker-dependent models is used as a quantitative measure of how much a given talker differs from another talker. An immediate application is talker identification: an unknown speaker may be identified by finding the closest speaker-dependent reference model to a model trained on the unknown speaker's data. Another application is to cluster similar talkers into a group; these may be used to train a HMM model that represents that talker better than a more general model. It is shown that using the model "nearest" a novel talker enhances the performance of a talker-independent speech recognition system.

1. A DISTANCE METRIC

The distance[1] metric is based on the relative entropy between discrete HMM output probability distributions. The object is to obtain a quantitative measure of how well a particular model represents a given talker. This can be used to select the most appropriate model to use for recognition, or the "closest" talker-dependent model for talker identification. Unlike similar applications of this metric [1, 2], the distance is not explicitly based on temporal features (HMM transition probabilities or frame lengths) and is thus independent of time.

1.1. Relative Entropy

Given two probability distributions $p(\cdot)$ and $q(\cdot)$ on the same discrete space X, the relative entropy (also called *information divergence* or the *Kullback-Leibler distance*) between the distributions is defined as [3]:

$$(p||q) \triangleq \sum_{x_i \in X} p(x_i) \log \frac{p(x_i)}{q(x_i)}. \quad (1)$$

Though this is not symmetric (in general $(p||q) \neq (q||p)$), a symmetric measure may be constructed by taking the mean of $(p||q)$ and $(q||p)$

$$D(p,q) \triangleq \frac{(p||q) + (q||p)}{2}. \quad (2)$$

[1] "Distance" is used here loosely, as the measures discussed may not be symmetric or satisfy the Triangle Inequality.

If the same vector quantizer (VQ) is used for two HMMs, the output probabilities are on the same discrete space, and thus the relative entropy between them may be computed. For the experiments described here, vector quantization is done using the MMI-constructed decision trees of section 1.2., (though a conventional nearest-neighbor VQ would probably also suffice).

1.2. Tree Output Probability Models for Speech

A decision tree is used to partition the feature space into a number of discrete regions (analogous to the Voronoi regions surrounding reference vectors in a vector quantizer). Unlike K-means reference vector estimation, the tree is grown in a supervised fashion. Each decision in the tree involves comparing one element of the vector with a fixed threshold, and going to the left or right child depending on the outcome. Each threshold is chosen to maximize the mutual information between the data and class labels (obtained from Viterbi alignment) that indicate the acoustic class of each datum. (The interested reader is referred to [4] for more information on tree construction.) The tree partitions the feature space into L non-overlapping regions or "cells," each of which corresponds to a leaf of the tree.

Let $p_j(i)$ denote the probability that an observation $\vec{o}_t$ emitted by HMM state s_i falls in leaf cell l_j (where i indicates the particular state and j the particular leaf):

$$p_j(i) = \Pr(l_j|s_i). \quad (3)$$

A tree density model therefore consists of a set P of probabilities $p_j(i)$, which can be considered as S vectors of length L, where S is the number of states in the model and L is the number of leaves in the tree. A tree-based HMM model can then be denoted as $\lambda \triangleq \lambda(\pi, A, P)$, where P takes the place of the output probability matrix B of conventional discrete HMMs. Note that if a tree is used as a vector quantizer, a model trained with quantized observations will have probabilities B identical to a tree-based model P. HMM parameters may be estimated from either the Baum-Welch algorithm as well as Viterbi training. The Viterbi algorithm is used for the experiments presented later and is discussed in the following section.

1.3. Tree Parameter Estimation

Given Viterbi-aligned data and an existing decision tree, it is straightforward to count NC_{ij}, the number of data points from state s_i (i.e. the number of observation vectors aligned with state s_i) that wind up in leaf l_j. From this, the conditional probability of an output falling in leaf node l_j given the state s_i may be estimated as

$$p_j(i) = \Pr(l_j|s_i) \quad (4)$$

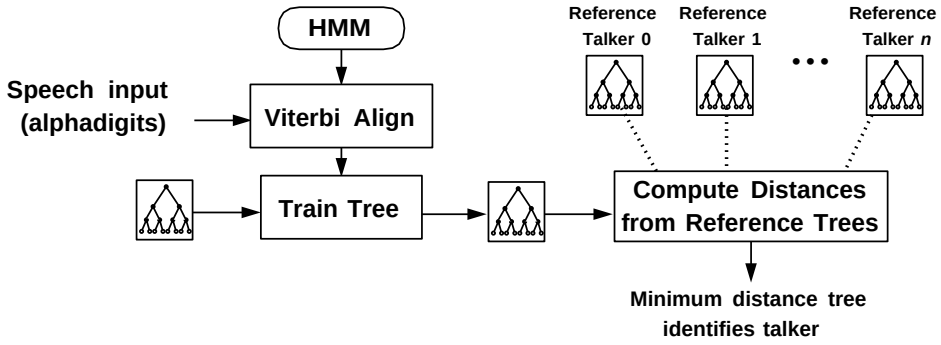

Figure 1. Talker Identification System

$$\approx \frac{NC_{ij}}{\sum_j NC_{ij}}. \quad (5)$$

Thus the tree output probabilities can be quickly estimated by Viterbi-aligning novel data and counting the relative frequencies at each leaf. (Alternatively, $p_j(i)$ may be estimated from the well-known forward-backward algorithm [4].) In practice, the talker identification task requires that models be trained with a minimum amount of data, leading to zero NC_{ij} for many i and j. In this case, $p_j(i)$ values of zero are set to some small floor value and renormalized.

Training identical trees with different data sets results in two (or more) different output probability estimates, denoted $p'_j(i)$ and $p''_j(i)$. The symmetric relative entropy between two distributions can then be computed by

$$D(p', p'') = \frac{\sum_i \sum_j \left[p'_j(i) \log \frac{p'_j(i)}{p''_j(i)} + p''_j(i) \log \frac{p''_j(i)}{p'_j(i)} \right]}{2}. \quad (6)$$

This is the distance metric used for the remaining discussion.

2. TREE-BASED QUANTIZATION FOR SPEECH

The tree-based vector quantizers have some interesting advantages over conventional vector quantizers. Perhaps most importantly, MMI-constructed trees can arguably handle the "curse of dimensionality" better than a minimum-distortion VQ, in part because only one dimension is considered at each split. Dimensions that do not help class discrimination are ignored, in contrast to a distortion metric which is always computed across all dimensions.

Another way to simplify high-dimensional problems is to use overlapping trees. In a manner similar to using multiple "codebooks," multiple trees may be grown independently for lower-dimensional subsets of the feature space. Output probabilities are then computed as the product of the tree probabilities. These will be underestimated if the subspaces are not truly independent, but this is usually outweighed by the higher spatial resolution obtainable (it increases exponentially with the number of overlapping trees).

Another advantage of trees is they are easily pruned to a smaller size, allowing the number of free parameters (proportional to the number of leaves) to be tailored to the problem. Where data is sparse (as in talker ID), smaller trees are more robust to undertraining, though the resulting model is coarser.

2.1. Experimental Tree Models

The feature space was seven adjacent 14-dimensional vectors consisting of 12 LPC-derived mel-cepstral coefficients, energy, and delta energy. This 98-dimensional space was divided into 5 subspaces consisting of cepstral coefficients 1–3, 4–6, 7–9, 10–12, and the energy/delta energy features, and separate trees (denoted 1–5 respectively) were constructed for each subspace. Thus the basic model consisted of 5 overlapping trees, one for each subspace. The trees used were derived from those that achieved the best performance (11.6% error) on a talker-independent connected alphadigit task [4]. (Because only seconds of data are available for the identification task, the trees were pruned from several thousand leaves to approximately 256 to reduce the number of parameters.)

Models were trained and tested using a corpus of connected alphadigits recorded at Brown University. Nearly 6.5 hours of speech from 116 talkers was collected using a head-mounted microphone sampled at 16kHz. Each talker read approximately 40 utterances composed of random digit sequences, random alphadigit sequences, and spellings of dictionary words. Utterances have an average length of about 15 connected alphadigits; these were truncated if the talker-identification task needed shorter test utterances. The database was arbitrarily split into a 96 talker (64 men, 32 women) training set, and an evaluation set of 20 different talkers (12 men, 8 women).

3. TALKER IDENTIFICATION

A good talker distance metric should be useful for talker identification, because models trained on data from the same talker will be "close." Figure 1 shows a talker identification scheme based on the distance measure of Section 1. Initially, a number of reference models are constructed, each trained on data from a particular speaker. A small amount of data from an unknown talker is Viterbi-aligned and used to train a "test" model. The distances between the test model and all reference models are computed; the reference model with the smallest distance from the test talker is assumed to identify the unknown talker. For speaker verification tasks, (i.e. the talker may not be in the training set), a threshold may be set to reject talkers that do not sufficiently resemble any reference model.

3.1. Identification Experiments

About 80 seconds of Viterbi-labeled speech data from each of the evaluation talkers were used to train 20 different "reference" models. For each talker, a "test" model was trained with 10 seconds of novel data not used for the reference model training. The relative-entropy distance was

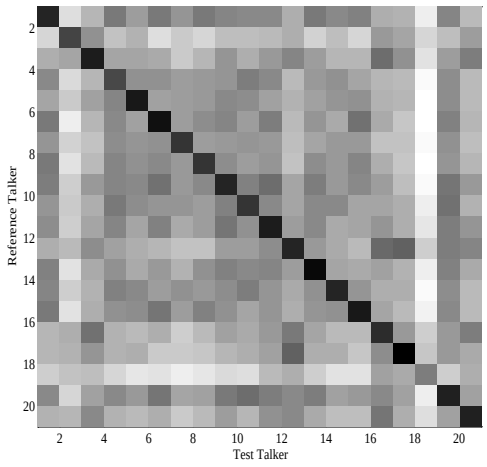

Figure 2. Talker distance matrix (darker = closer)

computed between each test and reference model; in all cases the distance to the same-talker reference model was smaller than any intratalker distance. Thus the distance metric used for talker identification results in 100% identification accuracy on this test set. Figure 2 displays the talker distance matrix graphically. The intersection of row i and column j represents the distance between the talker i reference model and the talker j test model; the lighter the element the more distant the talker. The closest distances in a row are clearly on the diagonals, which are the distances between the test and reference models from the same talkers.

A useful talker identification system should require only a small amount of test data. However, identification error will increase as the amount of test data shrinks, because many relative frequencies will be zero. To investigate the identification accuracy as a function of available test data, the same 20-talker experiment was performed but the amount of data available for the test models was varied from three to ten seconds, as shown in Figure 3. Note that six seconds of test data was sufficient to accurately identify all talkers. More phonetically balanced test utterances, would probably reduce the amount of necessary speech even further.

An additional experiment judged the accuracy of the talker identification over the 96-talker training database. Results with 10 seconds of test speech showed that only one talker was misidentified, a 1.04% error rate. Using 40 seconds of test data improved the accuracy to 100%, although this probably could have been achieved with substantially less than 40 seconds. Table 1 shows the results of using the distance measures of individual trees as well as the mean distance across all trees. The energy features (tree 5) were the least useful for talker identification, while the the higher-order cepstra (tree 4) gave the best performance alone. This is reasonable because energy and gross spectral shape are probably less talker-dependent than the more detailed spectral features represented by the the high-order cepstra. All features, however, were necessary for the highest identification rate.

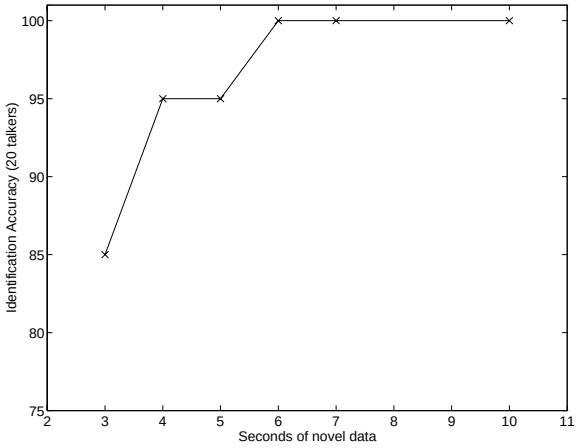

Figure 3. Talker Identification Accuracy vs. Test Speech

4. TALKER CLUSTERING FOR IMPROVED HMM RECOGNITION

It is generally agreed that using separate models for male and female talkers improves the performance of HMM-based speech recognition systems [5]. Unfortunately, often female-trained models are better for some male speakers and vice-versa. One use of the the distance measure is to aggregate talkers based on speech characteristics rather than gender. To demonstrate this, a partitioning algorithm ([6]) was employed to generate two different clusters of the 96 talkers in the training set. Half of the training talkers were randomly assigned to each cluster, then an iterative method was used to refine the clusters. A talker was moved from one cluster to the other if it increased a measure of cluster "goodness." The iteration was terminated when moving any talker resulted in a decrease of cluster goodness.

There are several reasonable ways to measure cluster goodness, including mean intracluster distance, mean intercluster distance, or a combination. For this experiment, the the mean intracluster talker distance, summed across the clusters, was used. This was computed by averaging all distances $D(i, j)$ for talkers i and j in the same cluster. After the iterative clustering, 90% of talkers in one cluster were male and 100% of talkers in the other cluster were female; this is good evidence that the clustering procedure groups similar talkers. The clustering is illustrated in Figure 4, which shows the distance matrix sorted by cluster (which are clearly visible). Note that one cluster is much larger (71 talkers) than the other (25 talkers); this is almost certainly due to the numerical disparity of females (32) and males (64) in the training set.

4.1. Recognition Experiments

Two different models were trained from the talker clusters just found. Recognition experiments were performed using the LEMS talker-independent, connected-alphadigit recognition system [7].) A further application of the distance metric is the selection of an appropriate model for a given talker. Figure 5 shows that the most appropriate model is

Tree	1	2	3	4	5	Mean
Error	8.3%	10.4%	8.3%	4.1%	24.0%	1.0%

Table 1. ID error for single and multiple trees.

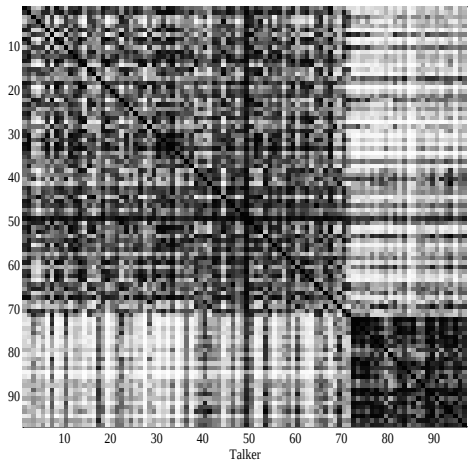

Figure 4. Distance matrix sorted by cluster

the one closest – that is, with the smallest distance – to a novel talker. The vertical axis is the difference in distances between talker-dependent "test" models and the two cluster models; a negative value means the talker was closest to the predominantly-male model. The horizontal "performance" axis is the difference in percentage word recognition error between the two models. A negative value means a given talker performed better on the predominantly-male model. The figure shows that for all talkers, recognition performance is substantially better on the "closest" model. Male talkers are indicated by "o" and female talkers by "+;" note that one female talker is best represented by the male model. In general, the distance metric is a better criterion than gender for selecting the appropriate model.

Using the cluster models improves recognition performance substantially. A 5-tree, 256-leaf baseline model was trained on all 96 talkers and resulted in 14.3% word error rate on the test set of 20 novel talkers. Additionally, two models were trained from the clustered talkers as just described. Using the closest model for recognition reduced the error from the baseline of 14.3% to 12.1%, a decrease of 15%. Perhaps more importantly, the word error of the poorest-performing talker was decreased by more than 40%, because the clustered model better represented that talker. Note that the comparison is between the baseline model trained on 96 talkers and models trained on 75 or 21 talkers, so even though the two models have twice as many parameters to estimate, the clustering partitions the data appropriately.

5. CONCLUSIONS

The work presented here could be extended in a number of ways. It should be emphasized that the trees were built to maximize the discrimination between different acoustic events and to effectively ignore talker-dependent features. The talker-identification distances might be made even more robust by constructing the trees to maximize discrimination between talkers. In this case, the class-independent output probabilities (the total probability mass in a leaf) could be used to obviate the need for the initial Viterbi alignment.

Baseline	By Gender	By Distance
14.3%	14.6%	12.1%

Table 2. Recognition error of multiple models.

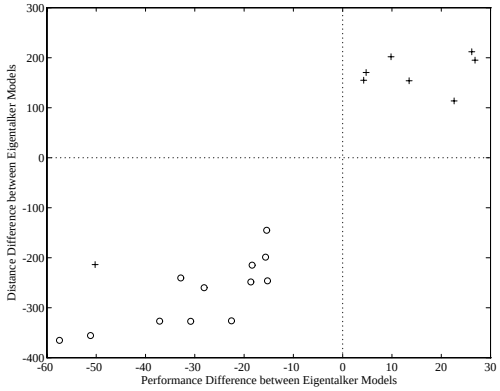

Figure 5. Model Distance vs. performance difference.

Clustered-talker modeling provides another area to explore. Different clustering algorithms and goodness-of-cluster measures should probably investigated. Given sufficient training data, more than two models could certainly be used; the optimal number of clusters is a open question. There is also no good reason to train separate models on distinct sets of talkers. Indeed, some overlap might be desirable to increase the amount of training data in a cluster.

To summarize, the distance measure described here captures meaningful differences among talkers and is rapidly computable. Experiments have shown that it is a robust means of talker identification, and also and serves as a quantitative way of clustering talkers to decrease the error of an HMM recognition system.

This work has been funded in part by NSF grant MIP-9120843.

REFERENCES

[1] B. H. Juang and L. R. Rabiner. A probabilistic distance measure for hidden Markov models. *AT&T Tech. J.*, 64:391–408, February 1985.

[2] A. Higgens, L. Bahler, and J. Porter. Voice identification using nearest-neighbor distance measure. In *Proc. 1993 ICASSP*, volume II, pages 375–379, April 1993.

[3] T. Cover and J. Thomas. *Elements of information theory.* John Wiley & Sons, Inc., New York, 1991.

[4] Jonathan T. Foote. *Decision-Tree Probability Modeling for HMM Speech Recognition.* Ph.D. thesis, Brown University, Providence, RI, 1993.

[5] F. Kubala and R. Schwartz. A new paradigm for speaker-independent training. In *Proc. 1991 ICASSP*, pages 833–836, May 1991.

[6] B. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell Sys. Tech. J.*, pages 291–301, February 1970.

[7] M. Hochberg, J. Foote, and H. Silverman. The LEMS talker-independent connected speech alphadigit recognition system. Technical Report 82, LEMS, Division of Engineering, Brown University, Providence RI, 1991.