HETEROGENEOUS MEASUREMENTS AND MULTIPLE CLASSIFIERS FOR SPEECH RECOGNITION[1]

Andrew K. Halberstadt and James R. Glass

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA
{drew, jrg}@sls.lcs.mit.edu

ABSTRACT

This paper addresses the problem of acoustic phonetic modeling. First, heterogeneous acoustic measurements are chosen in order to maximize the acoustic-phonetic information extracted from the speech signal in preprocessing. Second, classifier systems are presented for successfully utilizing high-dimensional acoustic measurement spaces. The techniques used for achieving these two goals can be broadly categorized as hierarchical, committee-based, or a hybrid of these two. This paper presents committee-based and hybrid approaches. In context-independent classification and context-dependent recognition on the TIMIT core test set using 39 classes, the system achieved error rates of 18.3% and 24.4%, respectively. These error rates are the lowest we have seen reported on these tasks. In addition, experiments with a telephone-based weather information word recognition task led to word error rate reductions of 10–16%.

1. INTRODUCTION

The acoustic-phonetic modeling component of most current speech recognition systems calculates a small set of homogeneous frame-based measurements at a single, fixed time-frequency resolution. This paper presents a contrasting approach, using more detailed and more diverse acoustic measurements, which we refer to as heterogeneous measurements. Diverse measurements are obtained by varying the time-frequency resolution, the spectral representation, the choice of temporal basis vectors, and other aspects of the preprocessing of the speech waveform. Using a wide variety of measurements leads to high-dimensional acoustic measurement spaces. This presents a challenge because the amount of training data needed to train a classifier grows exponentially as the dimensionality increases. This potential difficulty is avoided by dividing the measurements into subsets and training a separate classifier for each subset of measurements. The problem is thus transformed into determining how to combine the outputs of multiple classifiers. In our previous work [5], we reported on hierarchical techniques for combining classifiers. This paper focuses on committee-based approaches including voting, linear combination, and using an independence assumption. Hybrid methods combining elements of hierarchical and committee-based approaches are also presented. Phonetic recognition and telephone-based word recognition experiments show that these techniques generalize well to a variety of tasks and acoustic environments.

[1]This material is based upon work supported by the National Science Foundation under Grant No. IRI-9618731.

2. TASKS, CORPORA, AND CLASSIFIERS

Phonetic classification and recognition experiments were conducted using the TIMIT acoustic-phonetic corpus [8]. In accordance with common practice [9], we collapsed the 61 TIMIT labels into 39 labels before scoring. Glottal stops were ignored for classification, but were retained for recognition. We used the standard NIST 462 speaker training set, and 24 speaker core test set for final testing. An independent set of 50 speakers was used for system development. Word recognition experiments were performed using the JUPITER telephone-based weather information task [4]. Mixture diagonal Gaussian classifiers were used in all experiments. Normalization and principal components analysis were performed to whiten the feature space. For TIMIT, the segment models described below used a minimum of 61 data points per mixture component and a maximum of 96 mixtures per phone; the boundary models used a minimum of 10 data points per mixture component, and a maximum of 100 mixtures per linguistic unit. Model aggregation [6] of 4 training trials per classifier was used for all TIMIT experiments to improve the performance and robustness of the models. For JUPITER, the classifier used a minimum of 50 data points per mixture component, and a maximum of 50 mixtures per linguistic unit. Aggregation was not used for JUPITER.

3. HETEROGENEOUS MEASUREMENTS

We divide acoustic measurements into segmental measurements, which are calculated based on a start and end time, and "boundary", or landmark, measurements which are calculated using a single time specification. Figure 1 summarizes the characteristics of eight segmental (S1–S8) and five boundary (B1–B5) measurements used in subsequent experiments. In all measurements, a frame rate of 200 frames per second (5 ms per frame) was used for short-time Fourier transform (STFT) analysis. The first column is a label for ease of reference. The second column indicates the number of dimensions in the measurement set. For B1 and B2, the notation $104 \Rightarrow 60$ indicates that principal components analysis was used to reduce the dimensionality of the measurements from 104 to 60. The third column indicates the duration in milliseconds of the Hamming window for short-time Fourier transform analysis. The fourth column includes the spectral representation, which may include MFCCs or PLPCCs, energy, low frequency energy (LFE), and/or zero crossing (ZC) rate. The fifth column indicates the temporal basis that was applied. In each case, the temporal basis was applied as an inner product with the frame-based spectral representation. For the segmental

	# Dims	STFT [ms]	Spectral Representation	Temporal Basis
S1	61	10	12 MFCC	5 avg
S2	61	30	12 MFCC	5 avg
S3	61	10	12 MFCC	5 cos $\pm$ 30ms
S4	61	30	12 MFCC	5 cos $\pm$ 30ms
S5	64	10	9 MFCC	7 cos $\pm$ 30ms
S6	61	30	15 MFCC	4 cos $\pm$ 30ms
S7	61	20	12 PLPCC	5 avg
S8	61	20	12 PLPCC	5 cos
B1	104 $\Rightarrow$ 60	30	12 MFCC+ energy	8 avg 5 10 20 40
B2	104 $\Rightarrow$ 60	20	12 PLPCC+ energy	8 avg 5 10 20 40
B3	60	30	12 MFCC	5 cos $\pm$ 75ms
B4	60	30	12 MFCC+ZC+ energy+LFE	4 cos $\pm$ 50ms
B5	60	10	10 MFCC	6 avg 20 20 20

Table 1: Segmental and boundary measurement set summary.

measurements, the cosine temporal basis extends 30 ms beyond the start and end of the segment on both sides, indicated by $\pm$30. For the boundary measurements, the cosine basis extended 50 or 75 ms to either side of the boundary. For segmental measurements, the "5 avg" basis consists of averages over the segment in a 3-4-3 proportion, and also includes a 30 ms average on either side of the segment. For the boundary measurements, the "8 avg" basis consists of symmetric, non-overlapping averages over 5, 10, 20, and 40 milliseconds (indicated by 5 10 20 40) [3], for a total extension of 75 ms to either side of the boundary. The width of the average is increasing as the distance from the boundary increases. Similarly, the "6 avg" basis consists of symmetric, non-overlapping averages over 20, 20, and 20 milliseconds, for a total extension of 60 ms to either side of the boundary.

4. COMMITTEE-BASED METHODS

In this section, three methods are described for committee-based combination of classifiers. Let $A = \{\alpha_1, \alpha_2, \ldots\}$ be an ordered set of linguistic labels. Consider N classifiers which have been trained to discriminate among the elements of A. These classifiers may, in general, be defined over different measurement input spaces. Thus, for each input token and each classifier, n, there is a corresponding vector of measurements we denote by $\vec{f_n}$. For each token, let $\vec{f}$ be the column vector containing all of the measurements, that is, $\vec{f} = [\vec{f_1}^T \ \vec{f_2}^T \ \ldots \vec{f_N}^T]^T$, where ^{T} denotes the transpose operator. For each $\alpha \in A$ and each classifier, let $p_n(\vec{f_n}|\alpha)$ be the scalar value of the conditional probability density function (pdf) of $\vec{f_n}$. For each input token, the output of the acoustic modeling system is a vector of scores $\vec{s}$ with one entry for each linguistic unit.

The first method is to combine classifiers using voting. The output is the vector of scores from one of the individual classifiers. Ties are resolved by defining an ordering of the classifiers.

The second method is to combine scores linearly. In this technique, which we refer to as weighted linear combination of likelihood ratios (WLCLR), a likelihood ratio is used to normalize the absolute magnitude of the pdf values across classifiers. The equation for the scores is

$$s(\alpha_k) = \sum_{n=1}^{N} g_{k,n} \left(\frac{p_n(\vec{f_n}|\alpha_k)}{\sum_{a \in A} p_n(\vec{f_n}|a)} \right),$$

where the weights $g_{k,n}$ have the property $\sum_{n=1}^{N} g_{k,n} = 1 \quad \forall k$. Thus, the weights may be classifier specific and/or linguistic unit specific. All of the experiments reported here use equal weights. Alternatively, however, weights could be trained on a development set using a Maximum Likelihood (ML), Linear Least-Square Error (LLSE), or other criterion in order to automatically learn appropriate weights.

The third method is to combine classifiers by assuming statistical independence among the N random vectors $\vec{f_1}, \vec{f_2}, \ldots \vec{f_N}$, which leads to the expression

$$s(\alpha_k) = \prod_{n=1}^{N} p_n(\vec{f_n}|\alpha_k).$$

The feature vectors $\vec{f_1}, \vec{f_2}, \ldots \vec{f_N}$ in our experiments will seriously violate the independence assumption. In spite of this, empirical results demonstrate that this algorithm is an effective method for combining the outputs of multiple classifiers.

5. EXPERIMENTAL RESULTS

5.1. Comparing Committee-based Methods

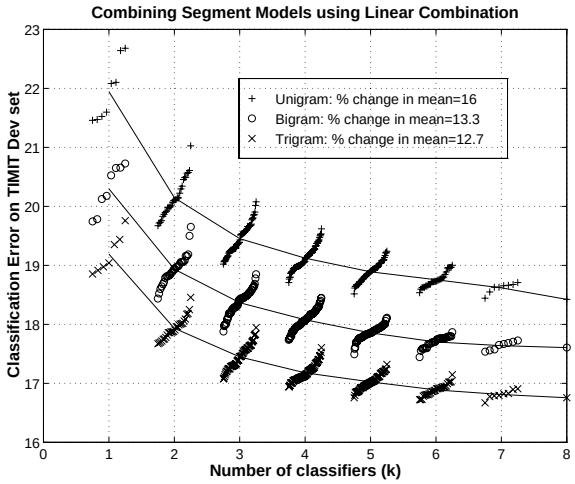

Figure 1: Combining classifiers using Linear Combination.

We compare the results of using voting, linear combination, or an independence assumption for combining multiple classifiers in the task of TIMIT phonetic classification. Rather than testing only a few configurations, Figures 1 and 2 show the performance of all possible subsets of the eight segmental measurements S1–S8. These figures show error rate versus k, the number of classifiers in the subset. For better viewing, individual data points are evenly spaced along the x-axis in the vicinity of the appropriate value of k. Lines connect the mean values. The total number of experiments for each phonotactic model is the sum for k equal to 1 through 8 of "8 choose k", which is 255. Unigram, bigram, and

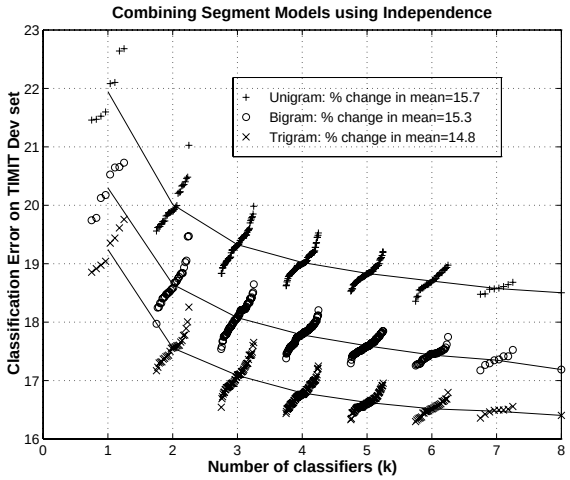

Figure 2: Combining classifiers using Independence.

trigram phonotactic models were used. The results indicate that indirect learning of phonotactic information has very little effect, since using multiple classifiers improves phonetic discrimination regardless of which phonotactic model is used. In addition to Figures 1 and 2, Table 2 summarizes some of the best CI classification results.

In the unigram case the three methods of voting, WLCLR, and independence produce similar performance. In contrast, in the bigram case, voting with 8 classifiers obtained 18.9%, which is actually worse than the 18.6% that was obtained with voting in the unigram case. This is probably because voting lacks soft-decision capability, and thus it does not improve the quality of the entire vector of scores, but rather focuses only on the quality of the top choice. The WLCLR and independence methods produce favorable trends with all three phonotactic models, although the independence assumption performs slightly better on average. In addition, the independence method is less expensive to implement, since the log scores can simply be added together, and it does not require calculation of a likelihood ratio. For these reasons, the remaining experiments with hybrid techniques, phonetic recognition, and word recognition all use the independence assumption to combine committees of classifiers.

Consider a pairwise distance metric between classifiers as the number of tokens which they classify differently on the dev set. Now generalize this metric to N classifiers by adding the pairwise distance between all classifiers in the set. We found that this generalized distance metric was correlated with the combined classifier performance. We observed correlation coefficients with magnitudes in the range of 0.45 to 0.63. Higher distance metrics led to lower error rates. Thus, given a set of classifiers, this metric can be used to predict which classifier combinations are likely to perform well.

5.2. Hierarchy/Committee Hybrids

In [5], we presented a MAP hierarchical approach to combining multiple classifiers. In this work, we have implemented two ways to combine hierarchical and committee-based approaches. The first hybrid approach uses a committee of classifiers at each node of a hierarchical tree. We implemented a hierarchical classifier as in [5], which uses different measurements for different

Methods	% Error Dev	core
Voting (S1–S8)	18.6	–
Linear Combination (S1–S8)	18.4	–
Independence (S1-S8)	18.5	–
Hybrid: Committees at nodes of tree	18.3	–
Hybrid: S1–S8 + Hierarchy	18.2	18.3

Table 2: Summary of TIMIT CI classification results.

Acoustic Measurements	Core Test set % Error	% Sub
avg of 1 seg + antiphone	30.1	19.0
avg of 1 seg + near-miss	28.7	18.0
5 segs + antiphone	27.7	16.7
avg of 1 bound	27.1	16.5
5 segs + near-miss	26.4	16.1
5 bounds	24.9	14.9
5 segs + 5 bounds + near-miss	24.8	15.0
5 segs + 5 bounds + antiphone	24.4	14.7

Table 3: Summary of TIMIT phonetic recognition results.

phonetic classes. Let us refer to the class-specific hierarchical measurement sets as SV, SN, SF, and SS, representing segmental vowel, nasal, fricative, and stop measurements, respectively. For example, we formed a hierarchy-of-committees classifier using: S1, S2, S4, and S5 at the root node; SV, S1, S4, S6, and S8 at the vowel node; SN, S2, S3, and S4, at the nasal node; SF, S1, S2, and S3 at the fricative/closure node; and SS, S1, S2, S5, and S8 at the stop node. Each of the committees was combined using an independence assumption. This resulted in a performance of 18.3% on the development set, as shown in Table 2. This hierarchical configuration suggests that computation can be reduced with minimal degradation in performance by targeting the measurements toward particular phone classes.

The second hybrid approach is to use a hierarchical classifier as one member of a committee. An implementation of the hierarchical classifier from [5] was added as a ninth member to the previously 8-member segmental measurements committee. The 9 classifiers were combined using independence to obtain 18.2% on the dev set, and 18.3% on the core set. This result is a 12.9% improvement over our previous best reported result of 21.0% [5]. The next best result that we have seen in the literature reporting TIMIT CI classification on the core test set is 23.0% [14].

5.3. Phonetic Recognition

Our TIMIT phonetic recognition experiments make use of a segment network produced by a first-pass recognition system. We refer to this step as probabilistic segmentation [1, 2, 10]. Either antiphone modeling [3] or 1-state near-miss modeling [1, 2] was used with segment models in order to account for both on-path and off-path segments in the segment network. All the phonetic recognition results make use of a phone bigram with a perplexity of 15.8 on the core set.

Table 3 summarizes a series of phonetic recognition experiments. The acoustic features for these experiments were S1, S2, S4, S5, S7 and B1–B5 from Table 1. The “avg of 1 seg” and “avg of 1 bound” rows refer to the average performance

Method	% Error core
Triphone CDHMM [7]	27.1
Recurrent NN [13]	26.1
Bayesian Triphone HMM [12]	25.6
Near-miss [2]	25.5
Heterogeneous Measurements	24.4

Table 4: Phonetic recognition results on TIMIT core set.

Acoustic Measurements	% Error	% Sub
B1	11.3	6.4
B4	12.0	6.7
B3 (altered)	12.1	6.9
3 bounds: B1 + B4 + B3(alt)	10.1	5.5

Table 5: Summary of JUPITER word recognition results.

over 5 experiments where each measurement set was used by itself. For the segmental performance, we report cases of using both near-miss modeling and antiphone modeling. When using the antiphone, combining 5 segmental measurements reduced the error rate from 30.1% to 27.7%, which is a 7.9% reduction. This change in performance is smaller than what was observed in classification. However, the substitution error fell from 19.0% to 16.7%, which is a 12.1% reduction. Thus, combining multiple classifiers has a significant effect in reducing substitution errors. Combining 5 boundary measurements reduced the error rate from 27.1% to 24.9%, and substitution errors fell from 16.5% to 14.9%. Adding segment models to the boundary models did not produce much additional gain, probably because the segment models were context independent (CI), while the boundary models were context dependent (CD). Near-miss models were better than the antiphone when using only segment models, but were worse when using segment and boundary models together. The final phonetic recognition result of 24.4% compares favorably with results in the literature. Table 4 compares this result with the best results reported in the literature.

5.4. Telephone-bandwidth Word Recognition

Finally, in order to verify that these techniques generalize to word recognition, we performed experiments using the telephone-based JUPITER weather information server task [4]. This particular configuration used an 1893-word vocabulary and a class bigram language model with a perplexity of 15.2 on the 1806 utterances in the test set. We trained three sets of boundary acoustic models (see Table 1), corresponding to B1, B4, and a variation of B3 with the STFT analysis window changed to 10 ms. Table 5 summarizes the results. Combining three boundary models led to error rate reductions of 10–16%, and substitution error rate reductions of 14–20%. These results confirm that these techniques generalize well to word recognition in a telephone bandwidth acoustic environment.

6. CONCLUSIONS

We have shown that heterogeneous measurements can be used to increase the acoustic-phonetic information extracted from the speech signal, and that combining multiple classifiers is an effective way to harness the discriminative ability of high-dimensional acoustic spaces.

This work has not been concerned about computational cost. In fact, using N different measurement sets increases the acoustic modeling cost by a factor of N in both memory and computation. Future work could consider how to obtain similar performance improvements at a much lower computational cost.

There is still a large gap between human and machine speech recognition ability, and current speech recognition systems rely more heavily on language models than humans do [11]. Once low-level acoustic-phonetic information is blurred or lost, it cannot be regained by subsequent processing, although the loss of acoustic-phonetic information may be masked by the application of higher-level lexical and linguistic constraints. Consideration of the results in this paper in light of the fundamental limits on time-frequency resolution and the non-invertibility of many preprocessing algorithms suggests that speech recognition systems of the future will need to include diverse acoustic measurements. In this way, more acoustic-phonetic information will be retained, the dependence on statistical language modeling will be decreased, and the gap between human and machine speech recognition performance will be narrowed.

7. REFERENCES

1. J. Chang and J. Glass, "Segmentation and modeling in segment-based recognition," in *Proc. EUROSPEECH*, pp. 1199–1202, 1997.
2. J. Chang, *Near-Miss Modeling: A Segment-Based Approach to Speech Recognition*. Ph.D. thesis, EECS, MIT, 1998.
3. J. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-based speech recognition," in *Proc. ICSLP*, Philadelphia, pp. 2277–2280, 1996.
4. J. Glass and T. Hazen, "Telephone-based conversational speech recognition in the jupiter domain," in *Proc. ICSLP*, Sydney, Australia, 1998. These Proceedings. To appear in 1998.
5. A. Halberstadt and J. Glass, "Heterogeneous measurements for phonetic classification," in *Proc. EUROSPEECH*, pp. 401–404, 1997.
6. T. J. Hazen and A. K. Halberstadt, "Using aggregation to improve the performance of mixture Gaussian acoustic models," in *Proc. ICASSP*, pp. 653–656, 1998.
7. L. Lamel and J. Gauvian, "High performance speaker-independent phone recognition using CDHMM," in *Proc. EUROSPEECH*, pp. 121–124, 1993.
8. L. Lamel, R. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Proc. of the DARPA Speech Recognition Workshop*, Palo Alto, February 1986. Report No. SAIC-86/1546.
9. K. Lee and H. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 11, pp. 1641–1648, November 1989.
10. S. Lee and J. Glass, "Probabilistic segmentation for segment-based speech recognition," in *Proc. ICSLP*, 1998. These Proceedings.
11. R. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, July 1997.
12. J. Ming and F. J. Smith, "Improved phone recognition using Bayesian triphone models," in *Proc. ICASSP*, pp. I–409–412, 1998.
13. A. Robinson, "An application of recurrent neural nets to phone probability estimation," *Transactions on Neural Networks*, vol. 5, pp. 298–305, March 1994.
14. S. Zahorian, P. Silsbee, and X. Wang, "Phone classification with segmental features and a binary-pair partitioned neural network classifier," in *Proc. ICASSP*, pp. 1011–1014, 1997.