HETEROGENEOUS ACOUSTIC MEASUREMENTS FOR PHONETIC CLASSIFICATION[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

In this paper we describe our recent efforts to improve acoustic-phonetic modeling by developing sets of heterogeneous, phone-class-specific measurements, and combining these diverse measurements into a probabilistic classification framework. We first describe a baseline classifier using homogeneous measurements. After comparing selected sub-tasks to known human performance, we define sets of phone-class-specific measurements which improve within-class classification performance. Subsequently, we combine these heterogeneous measurements into an overall context-independent classification framework. We report on a series of phonetic classification experiments using the TIMIT acoustic-phonetic corpus. Our overall framework achieves 79.0% accuracy on the NIST core test set.

1. INTRODUCTION

Over the past several years, our group has pursued a segment-based approach to speech recognition. One of the potential advantages of this approach over conventional frame-based methods is that it provides more flexibility in choosing *what* acoustic attributes to extract, and *where* to extract them from the speech signal. We believe such flexibility will be necessary to take full advantage of the acoustic-phonetic information encoded in the speech signal. To date we have used homogeneous feature vectors to represent the acoustic-phonetic information needed to discriminate among all sounds. Although this framework has worked well, obtaining good phonetic classification and recognition results [1, 4, 9], our recent work indicates that further gains can be obtained by incorporating heterogeneous, phone-class-specific measurements into our framework. There are at least two potential advantages to this approach. First, heterogeneous measurements provide the opportunity to develop diverse measurements which focus on the phonetically relevant information for discriminating among the sounds in a particular phone class. Second, heterogeneous measurements permit the removal of dimensions that are unnecessary in a particular phone class, thus making better use of the training data and reducing the computation required for classification.

There are two major challenges involved in implementing this strategy. First, measurement sets which provide within-phone-class improvements in phonetic classification must be developed. Second, the use of these heterogeneous measurements must be combined into an overall classification framework. We address both of these problems in this paper.

[1]This research was supported by DARPA under contract N66001-94-C-6040, monitored through Naval Command, Control and Ocean Surveillance Center.

2. TASK, CORPUS, AND CLASSIFIER

Unlike phonetic recognition, the task of phonetic classification makes use of an externally provided segmentation of the speech signal, thus eliminating any possibility of deletion or insertion errors. The classification experiments reported in this paper are context-independent (one model per phone), although the feature measurements were allowed to draw information from outside the boundaries of the current speech segment.

All experiments were conducted using the TIMIT acoustic-phonetic corpus [6]. In accordance with common practice [8], we collapsed the 61 TIMIT labels into 39 labels before scoring and we ignored glottal stops. For the experiments in this paper, we make reference to the manner classes of vowels/semivowels (VS), nasals/flaps (NF), stops (ST), weak fricatives (WF), strong fricatives (SF), and closures/silence (CL). Table 1 indicates the mapping to 39 phone classes and the manner label of each class.

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. For purposes of significance testing by McNemar's test [3], we also evaluated classification performance on a test set of 118 speakers, which was the full NIST 168 speaker test set minus our development set. Note that there is no overlap of speakers between any of the sets, and the training set has different sentences from the development and test sets. Table 2 indicates the number of tokens in the data sets on a class-by-class basis.

A mixture Gaussian classifier was used for all experiments, which made use of phone priors from the training data (e.g., a phone unigram). Normalization and principal component analysis were performed to whiten the feature space. For each trial of model training, a maximum of 12 full-covariance Gaussian kernels were allowed per phone. The mixture kernels were seeded via randomly initialized K-means clustering and trained using the EM algorithm. The number of mixtures was selected to achieve a minimum of approximately 500 tokens per kernel. Multiple trials of model training were performed and combined to produce more robust mixture models.

3. BASELINE CLASSIFIER

A baseline classifier was first established using homogeneous measurements. This measurement set reflects previous classification work [1] in our group.

After preemphasis and DC-offset removal, a short-time Fourier transform (STFT) analysis was performed every 5 ms using a 20.5 ms Hamming window. The STFT was converted to a set of 40 Mel-frequency spectral coefficients, which were then transformed to 12 Mel-frequency cepstral coefficients (MFCC's) using a cosine transform. A 61-dimensional homogeneous measurement vector was calculated for each phonetically labeled segment in the TIMIT transcriptions. The measurement vector consisted of three MFCC averages computed approximately over segment thirds (actually in a 3–4–3 proportion),

1	iy	VS	2	ih ix	VS	3	eh	VS
4	ae	VS	5	ah ax ax-h	VS	6	uw ux	VS
7	uh	VS	8	aa ao	VS	9	ey	VS
10	ay	VS	11	oy	VS	12	aw	VS
13	ow	VS	14	er axr	VS	15	l el	VS
16	r	VS	17	w	VS	18	y	VS
19	m em	NF	20	n en nx	NF	21	ng eng	NF
22	dx	NF	23	jh	SF	24	ch	SF
25	z	SF	26	s	SF	27	sh zh	SF
28	hh hv	WF	29	v	WF	30	f	WF
31	dh	WF	32	th	WF	33	b	ST
34	p	ST	35	d	ST	36	t	ST
37	g	ST	38	k	ST			
39	bcl pcl dcl tcl gcl kcl epi pau h#, CL							

Table 1: 39 phone classes from [8], and manner class membership.

Task	# of tokens in set			
	Train	Dev	Test	Core
Overall	140,225	15,057	35,697	7,215
Vowel/Semivowel	58,840	6,522	15,387	3,096
Nasal/Flap	14,176	1,502	3,566	731
Stop	16,134	1,685	4,022	799
Fric/Clos/Sil	51,075	5,348	12,722	2,589

Table 2: Number of tokens in each data set, ignoring glottal stops.

two MFCC derivatives computed over a time window of 40 ms centered at the segment beginning and end, and log duration.

This baseline configuration achieved classification accuracies of 78.9% and 78.4% on the development and core test sets, respectively. These results compare favorably with others previously reported in the literature [5, 9, 11, 13]. Figure 1 shows a bubble plot of the baseline classifier confusion matrix on the development set. Nearly 80% of the confusions occur by choosing an alternative in the correct manner class. Another 7% occur due to confusions involving the closure/silence class. This analysis suggested to us that if we could reduce the confusions within each manner class using class-specific measurements, then overall accuracy would also improve.

4. HUMAN VS MACHINE

Although we were able to achieve good performance on phonetic classification with our baseline classifier, we were interested to understand how machine performance compared with humans at the same task. For this purpose, we made use of previously reported perceptual studies concerning human classification performance using the TIMIT corpus. We examine the sub-tasks of vowel and stop classification below.

Cole and Methusamy [2] have performed perceptual studies on vowels excised from TIMIT. For this study, 16 vowel labels from TIMIT were selected, and 168 tokens of each were extracted, for a total of 2688 tokens. The 16 vowels for this study were /iy ih ey eh ae er ah ax aa ao uh uw ow aw ay oy/, or in IPA symbols, [iʸ ɪ eʸ ɛ æ ɝ ʌ ə ɑ ɔ ʊ uʷ oʷ ɑʷ ɑʸ oʸ]. The results indicated that vowels presented in isolation were identified with 54.8% accuracy, while vowels presented with one segment of context were identified with 65.9% accuracy. Our baseline classifier obtains 69.8% accuracy on the development set in this 16-way vowel identification task. Although the test sets are not exactly the same, this result indicates that humans and machines are performing about equally well in this task.

Lamel [7] has reported on perception of stop consonants extracted from TIMIT. The results are broken down according to context. We will consider three phonetic contexts, namely syllable-initial stops in a vowel-stop-vowel sequence, vowel-fricative-stop-vowel sequences, and non-syllable-initial stops in homorganic nasal clusters. Only short speech segments of three phones (V–S–V) or four phones (V–F–S–V and V–N–S–V) were presented to the listener. These sequences were sometimes across word boundaries, so listeners could not use lexical, syntactic, or semantic information. We obtained a list of the testing tokens so that we could use the same set. We trained our classifier on speakers that were not in the test set under consideration to ensure that the system remains speaker-independent. Our system was trained on stops from all contexts, so it provides a context-independent result.

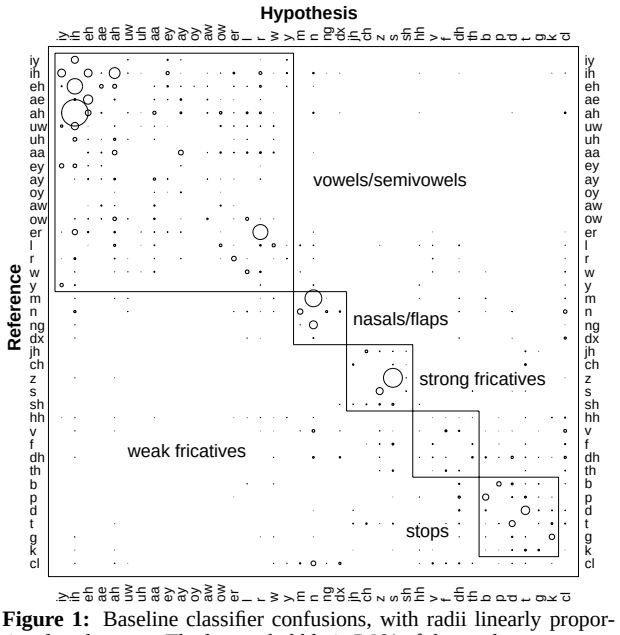

Figure 1: Baseline classifier confusions, with radii linearly proportional to the error. The largest bubble is 5.2% of the total error.

Phonetic Context	Human Error (%)	Machine Error (%)	Error Rate Increase Factor
V–S–V	3.4	11.3	3.3
V–F–S–V	12.2	32.9	2.7
V–N–S–V	7.6	18.7	2.5

Table 3: Human vs machine stop classification.

Table 3 summarizes the results. For syllable-initial singleton stops followed by a vowel, Lamel reports that human listeners achieved 3.4% error. The machine classifier performed more than three times worse, obtaining 11.3% error. For vowel-fricative-stop-vowel sequences, human listeners obtained 12.2% error, while the machine classification performed more than two and a half times worse, obtaining 32.9% error. For non-syllable-initial stops in homorganic nasal clusters, human listeners obtained 7.6% error on TIMIT tokens, while the machine obtained 18.7% error. Thus, the machines performed about two and a half times worse.

These results indicated that for stop consonants, there is a significant amount of low-level acoustic phonetic information which the automatic classifier is not effectively extracting. This experimental outcome is consistent with results in the literature comparing human and machine performance in a variety of speech recognition tasks [10]. These results also motivated our attempts to extract more low-level acoustic-phonetic information from the speech signal through the use of heterogeneous measurements.

5. HETEROGENEOUS FEATURES

Analysis of our classification systems showed that the within-phone-class performance was dependent on the time-frequency resolution of the Fourier analysis. Furthermore, the optimal settings for individual phone-classes differed substantially. For example, experiments with an earlier system showed that stop classification was optimized with increased time resolution, whereas the nasal consonants preferred decreased time resolution, as indicated in Figure 2. Therefore, when a single Hamming window duration is chosen, it is a compromise among the conditions that are favorable in different phone classes. This observation provided evidence for the hypothesis that a heterogeneous feature space could offer improvements in overall classification accuracy. In addition, as we noted in Figure 1, most phonetic confusions occur within the correct manner class. As a result, we have chosen initially to determine heterogeneous measurements to improve within-manner-class classification accuracies. In the current work we combined the three manner classes of weak fricatives, strong fricatives, and closures into a single class. In the following paragraphs we describe phone-class-specific measurements and report within-class classification accuracies on the development set. We compare the performance of these measurements to the baseline and also report the McNemar significance level of the difference.

For vowel/semivowel measurements, we used 62 dimensions. The first 60 dimensions were calculated as in [13]. These involve calculation of MFCC-like frame-based measurements, followed by a cosine transform in the time dimension to encode the trajectories of the frame-based features. The use of a tapered, fixed length (300ms) window in the cosine transform results in capturing some contextual information which can be modeled in an unsupervised manner through the mixtures in the Gaussian models. In addition to these 60 measurements, duration and average pitch were also included for a total of 62 measurements. The pitch measurement was calculated using a cepstral-based method. These measurements resulted in a vowel/semi-vowel accuracy of 74.3% on the development set, which improves upon the 73.1% (0.02 significance level) obtained by the baseline system, and is competitive with previously reported results [11].

For nasals, baseline measurements were altered by changing the Hamming window duration to 28.5 ms and adding a measure of average pitch, giving a total of 62 measurements per segment. These nasal-optimized measurements achieved 85.2% on the development set, compared to 83.4% obtained by the baseline system (0.001 significance level).

In our stop classification experiments, we increased the time resolution by using a 10 ms Hamming window, and used a 50 dimensional feature vector, composed of MFCC averages over halves of the segment (24 dimensions), time-derivatives of the MFCC tracks at the segment boundaries and at the start of the previous segment (24 dimensions), a measure of low-frequency energy in the previous segment (1 dimension), and log duration. We found that averaging over halves of the segment instead of thirds did not cause a drop in performance for the stops. Due to smaller dimensionality (50 dimensions), we adjusted the classifier by lowering the minimum number of tokens per Gaussian kernel from 500 to 300. In a six-way stop classification task, these measurements obtained 83.4% on the development set, compared to 79.6% for the baseline (10^{-4} significance level), and compare favorably to previously reported results [12].

For fricatives and closures, a 26.5 ms Hamming window was used for frame-based calculations. Time derivatives of only 11 MFCC's (instead of 12) were extracted at the segment boundaries. Three new measurements were added: the zero-crossing rate, the total energy of the entire segment (which is similar but not the same as the information in the first MFCC coefficient), and a time derivative of the low frequency energy at the beginning of the segment. This 62-dimensional measurement set obtained 91.2% on the development set, compared to 90.9% for the baseline (0.1 significance level).

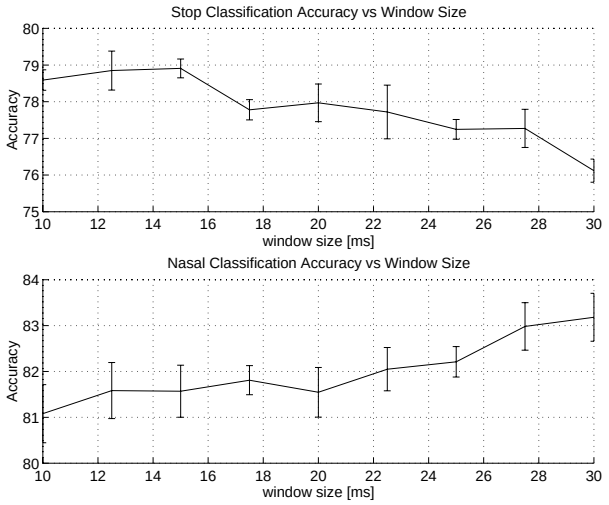

Figure 2: Comparison of within-class stop and nasal classification as a function of Hamming window duration. The vertical bars show one standard deviation in the performance calculated over five trials of mixture model training.

6. OVERALL FRAMEWORKS

The second major challenge which must be addressed in order to use heterogeneous measurements is to define a framework for overall classification which makes use of these diverse measurements. The goal of phonetic classification is to determine the most probable phone, α^*, given the acoustic feature vector $\mathbf{f}$. We can expand the decoding procedure over a set of phone classes C_i according to the expression

$$\alpha^* = \arg\max_{\alpha} P(\alpha|\mathbf{f}) = \arg\max_{\alpha} \sum_i P(\alpha|C_i,\mathbf{f})P(C_i|\mathbf{f}).$$

If each phone belongs to only one class, as is the case in this paper, then the summation over i becomes trivial since for each phone there is only one i such that $P(\alpha|C_i,\mathbf{f})$ is nonzero.

In these expressions, $\mathbf{f}$ represents all of the measurements that might be used by the system. Thus, each set of heterogeneous measurements is a subset of $\mathbf{f}$. In fact, we can cast the above decoding as a hierarchical process [1]. Thus, at level zero, a single measurement set $\mathbf{f}^{(0)} \subset \mathbf{f}$ is used to determine the probability of membership in class j at level one, that is

$$P(C_j^{(1)}|\mathbf{f}) \approx P(C_j^{(1)}|\mathbf{f}^{(0)}).$$

In this expression we have decided to approximate $P(C_j^{(1)}|\mathbf{f})$ by $P(C_j^{(1)}|\mathbf{f}^{(0)})$ based on practical considerations, such as problems with high classifier dimensionality and superfluous measurement dimensions. These considerations led us to the assumption that each class probability can be more accurately estimated in practice using a subset of the features contained in $\mathbf{f}$. This assumption does not necessarily hold from a purely theoretical standpoint, where issues stemming from finite training data can be ignored. Continuing at level one, the feature set used to determine the conditional probability of level two class membership can depend upon the conditioning level one class index, j. We indicate the feature set dependence on j using the notation $\mathbf{f}_j^{(1)}$. Thus the conditional probability of level two class membership is obtained using the approximation

$$P(C_i^{(2)}|C_j^{(1)},\mathbf{f}) \approx P(C_i^{(2)}|C_j^{(1)},\mathbf{f}_j^{(1)}).$$

Using this notation and the above approximations, our previous decoding equation becomes

$$\arg\max_i \sum_j P(C_i^{(2)}|C_j^{(1)}, \mathbf{f}_j^{(1)}) P(C_j^{(1)}|\mathbf{f}^{(0)}).$$

This process can be iterated to as many stages as desired. In the present implementation, the level two classes $C_i^{(2)}$ are the individual phones, so no further iteration is required. This MAP framework for combining heterogeneous measurements achieved 80.0% on the development set compared to 78.9 for the baseline (10^{-5} significance level), and was also used for final testing on the NIST core set to obtain 79.0%. When compared to the NIST core baseline of 78.4%, the significance level was 0.16. However, we suspected that the small size of the core set made significance testing somewhat coarse. Therefore, we also compared the baseline and heterogeneous framework results on the 118 speaker test set, which includes all data not in the training or development sets, with results summarized in Table 4. The overall results of 78.4% and 79.0% were the same as for the core set, but with better significance levels. These results confirm that heterogeneous measurements are producing significant improvements on independent testing data.

The above MAP framework allows for some interaction between the scores at different levels. Alternatively, we have implemented a strict framework in which the first classifier makes a hard decision about the level one class membership. This strict framework also achieved 80.0% on the development set, and fewer than 1% of the testing tokens were classified differently from the MAP framework. The strict framework requires the computation of only one level one feature set $\mathbf{f}_j^{(1)}$ for each segment, which provides an opportunity for computational savings compared to the MAP framework. This strict framework can be thought of as a strategy for pruning the full MAP framework, and other pruning strategies could also be devised which save computation with minimal effect on performance [1].

7. CONCLUSIONS

These experiments demonstrate the viability of using heterogeneous, phone-class-specific measurements to improve the performance of acoustic-phonetic modeling techniques. We have obtained preliminary solutions for handling the two challenges of developing diverse features to improve classification accuracy and having a framework to combine the features into an overall system. Our final results of 79.0% on the core test set compare favorably to results in the literature. Zahorian [13] reports 77.0% on the core test set, while Leung et al. [9] report 78.0% on a different test set.

The design of a feature extraction mechanism for pattern classification frequently leads to a tradeoff between retaining as much relevant information as possible while at the same time avoiding unmanageably high classifier dimensionality. Our studies of human/machine comparisons make us believe that conventional measurement extraction procedures err on the side of ignoring relevant phonetic information in order to streamline the classifier. The heterogeneous frameworks proposed in this paper provide an alternative way to deal with this tradeoff. In fact, the final classifier successfully manages to make use of 290 different measurements for each proposed segment. In addition, it is important to recognize that conventional measurement extraction procedures which are optimized over all phones tend to settle on a compromise among the measurements that are best in different phone classes. It is not clear if heterogeneous features will ultimately outperform homogeneous features for typical speech recognition tasks. For example, if the modeling difficulties and computational burden could be overcome, it is likely that using all the measurements proposed in this paper jointly in a single classifier would match or exceed the performance achieved here using a hierarchy of multiple classifiers. However, this line of investigation has generally not been taken because of the difficulties which surround it.

Task	Baseline	Heterogeneous	Significance
Overall	78.4	79.0	0.001
Vowel/Semivowel	72.2	72.7	0.18
Nasal/Flap	83.5	84.6	0.004
Stop	80.4	82.1	0.002
Fric/Clos/Sil	90.9	91.2	0.06

Table 4: Classification accuracies on the 118 speaker test set.

In future work, we plan to search for other sets of measurements which improve within-class classification performance, and to use different partitions of the phones into classes. We would like to explore alternative methods of combining heterogeneous measurements and multiple classifiers. In addition, we plan to develop techniques for using these phone-class-specific measurements in phonetic recognition and word recognition.

We would like to thank Lori Lamel for providing us with token lists from her stop perception experiments. We would like to thank T.J. Hazen and Kenney Ng for ideas and discussion regarding robust training procedures.

8. REFERENCES

[1] R. T. Chun. *A Hierarchical Feature Representation for Phonetic Classification*. M.Eng. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, March 1996.

[2] R. A. Cole and Y. K. Methusamy. Perceptual studies on vowels excised from continuous speech. In *ICSLP*, pages 1091–1094, Banff, Canada, October 1992.

[3] L. Gillick and S.J. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *ICASSP*, pages 532–535, Glasgow, Scotland, May 1989.

[4] J. Glass, J. Chang, and M. McCandless. A probabilistic framework for feature-based speech recognition. In *ICSLP*, pages 2277–2280, Philadelphia, October 1996.

[5] W. D. Goldenthal. *Statistical Trajectory models for Phonetic Recognition*. Ph.D. thesis, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, September 1994.

[6] 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.

[7] L. F. Lamel. *Formalizing Knowledge used in Spectrogram Reading: Acoustic and Perceptual Evidence from Stops*. Ph.D. thesis, Department of Electrical and Computer Engineering, Massachusetts Institute of Technology, Cambridge, May 1988.

[8] K. F. Lee and H. W. Hon. Speaker-independent phone recognition using hidden Markov models. *IEEE Trans. Acoust., Speech, Signal Processing*, 37(11):1641–1648, November 1989.

[9] H. Leung, B. Chigier, and J. Glass. A comparative study of signal representations and classification techniques for speech recognition. In *ICASSP*, pages 680–683, Minneapolis, April 1993.

[10] R. P. Lippmann. Speech perception by humans and machines. In *Proc. of the ESCA Workshop on the "Auditory Basis of Speech Perception"*, pages 309–316, Keele University, U. K., July 1996.

[11] P. Schmid. *Explicit N-best Formant Features for Segment-Based Speech Recognition*. Ph.D. thesis, Department of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, Portland, October 1996.

[12] X. Wang, S. A. Zahorian, and S. Auberg. Analysis of speech segments using variable spectral/temporal resolution. In *ICSLP*, pages 1221–1224, Philadelphia, October 1996.

[13] S. A. Zahorian, P. Silsbee, and X. Wang. Phone classification with segmental features and a binary-pair partitioned neural network classifier. In *ICASSP*, pages 1011–1014, Munich, Germany, April 1997.