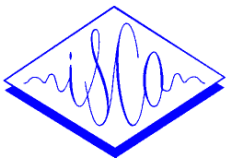

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

MODELLING SPECTRAL DYNAMICS FOR VOWEL CLASSIFICATION[1]

William D. Goldenthal and James R. Glass

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA
email: thal & jrg @goldilocks.lcs.mit.edu

ABSTRACT

In this work, we are attempting to develop models which capture the dynamic characteristics and statistical dependencies of acoustic attributes in a segment-based framework. Our approach is based on the creation of a *track*, $\vec{T}_\alpha$, for each phonetic unit α. The track serves as a model of the dynamic trajectories of the acoustic attributes over the segment. The tracks attempt to capture segment-level spectral dynamics without making any assumptions concerning the linearity or stationarity of the speech signal. The statistical framework for scoring incorporates the auto- and cross-correlation properties of the track error over time, within a segment. This paper presents the results of a series of vowel classification experiments using the TIMIT acoustic-phonetic corpus. Classification performance of 68.9% was achieved, which compares favorably to other vowel classification experiments using the same corpus.

INTRODUCTION

The speech signal varies slowly enough that successive observation frames of acoustic attributes are highly correlated in time. However, the most successful speech recognition algorithm developed thus far (i.e. HMM's), models the speech signal as a sequence of observations whose probabilities are dependent on the state, and are independent of other observations in the utterance. Researchers have been successful in trying to implicitly incorporate contextual information into the observation vector by augmenting it with adjacent frames [1], or with first (and possibly second) order spectral differences [2, 3]. However, there is clearly additional correlation information present in the speech signal.

Segment-based approaches hypothesize start and end times for each segment during the matching process. While assuming independence between adjacent segments, these formulations offer the possibility of directly modelling the within-segment correlations. These approaches tend to fall into one of two categories. In one case an N frame speech segment, $S = \{\vec{s}_1, \ldots, \vec{s}_N\}$, is transformed to match a model of some fixed dimension. This formulation includes work which maps S into a fixed dimensional vector, $\vec{x}$, which is used as the basis for classification [4, 5, 6, 7, 8, 9, 10]. In these cases, the vector must contain all relevant dynamic information about the segment. This case also includes work which quantizes S into a fixed number of frames, M, and models the likelihood of the quantized segment [11].

[1]This research was supported by ARPA under Contract N00014-89-J-1332 monitored through the Office of Naval Research. W.D. Goldenthal receives support from C.S. Draper Laboratory.

The second common segment-based approach is one where each model generates an N frame synthetic segment G which is matched to S. Segment-based HMM's fall into this category [12, 13], as does work which attempts to parameterize the dynamics of the acoustic attributes over the course of the segment [14, 15, 16]. Many of these formulations resort to a frame-by-frame classification within the segment, thereby reducing computation, but potentially foregoing some of the benefits of a segment-based framework.

Our formulation falls into the second catagory. An N frame synthetic segment G, is generated and compared directly to S. An advantage of this approach is that all of the data is utilized. Classification is then performed at the segment level. Our goal is to model both the time-varying spectral dynamics within a segment, as well as the statistical dependencies. In the following section, we outline this approach in more detail, describing our method for generating a synthetic segment, as well as our error modelling methodology. This is followed by a description of vowel classification experiments using the TIMIT acoustic-phonetic corpus [17]. Finally, we discuss the results and outline our plans for extending this work.

MODELLING SPECTRAL DYNAMICS

The main emphasis of this work is the creation of a dynamic, phone dependent *track* which can be defined as a trajectory, or temporal evolution of the acoustic attributes over a segment. A track consists of a sequence of M state vectors $T = \{\vec{t}_1, \ldots, \vec{t}_M\}$ which are used as the basis for generating a synthetic segment

$$G = f(T, N) = \{\vec{g}_1, \ldots, \vec{g}_N\} \quad (1)$$

for any number of frames N, where $f()$ is a generation function. The track serves as a template, and attempts to capture segment-level spectral dynamics.

To classify an N frame speech segment, S, we generate synthetic segments, G for each phonetic model α. The match between S and G is made by computing the error

$$E = S - G = \{\vec{e}_1, \ldots, \vec{e}_N\} \quad (2)$$

where

$$\vec{e}_i = \vec{s}_i - \vec{g}_i \quad (3)$$

The likelihood of each phonetic model, α, is based on the error probability $P(E|\alpha)$. The statistical framework for scoring incorporates the auto- and cross-correlation properties of the

track error over time, within a segment. Hence, statistical dependencies are incorporated into the model.

There exist other approaches in the literature which attempt to explicitly model the segment level dynamics. Gish and Ng create trajectory models which assume the acoustic features vary as a first or second order polynomial [16]. Digilakis estimates the acoustic parameters based on propagating the states through a sequence of stochastic dynamical systems. The acoustics are assumed to be piece-wise linear and stationary [15].

In this work we do not parameterize the spectral trajectories. Instead, we create tracks from the data by mapping it to a sequence of M states for each phone. When all the tokens in the training set for a particular phone have been mapped, the phone dependent track is calculated from the ML estimate of each state. The objective is to capture the non-linear nature of the dynamics in a simple way, without explicitly assuming anything a-priori about the linearity or stationarity of the signal.

Regardless of the method used to create a track for the data, an important question that must be answered is how to generate a synthetic segment G of a particular number of frames. Accounting for durational variability is a key component of any segment-based dynamic model. Digilakis defined two terms known as *trajectory invariance* and *correlation invariance* [14], which occupy opposing ends of a continuum for accounting for the durational variability and generating the required alignment.

Trajectory invariance is based on the assumption that each phone's attributes follow a fixed trajectory through the acoustic space, and that the trajectory is sampled at a number of points equal to the observation length of the phone. A trajectory invariance model consists of a set of M states. Segments which are less than M frames are mapped to a subset of the track states. In cases where the segment is longer than M frames, the segment is either sub-sampled [11], or the track is expanded to the same number of frames as the input token [15].

The correlation invariance assumption is based on the idea that the correlation between measurements depends only on the relative position of the observations within a segment. The trajectory through the acoustic space varies with phone duration, and the correlation between successive frames is considered to be invariant. An example of this technique occurs in [15], where Kalman filters were used to propagate an estimate of the state for the duration of the segment.

In our initial attempts to select a method for creating a track, T, and its associated generation function, $f()$, we explored several different alternatives. A simple comparison was made by measuring the total distortion error on the training set between each token, S and its synthetic counterpart, G where the distortion for a particular token is defined by the mean square error

$$D = \frac{1}{N}\sum_{i=1}^{N} \|\vec{e}_i\| \qquad (4)$$

In order to make a Euclidean distance more reasonable, all dimensions were first normalized by their standard deviations.

Distortion measurements revealed that of the alternatives we investigated the trajectory invariance assumption was consistently slightly superior. The generation function which resulted in the least distortion was *fixed endpoints with interpolation*. Hence, the remainder of the paper is based on tracks created using this algorithm.

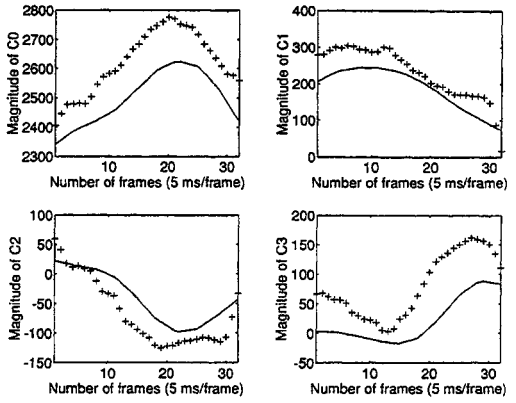

Figure 1: Mel-frequency cepstral coefficients C0-C3 for a synthetic [ɔʸ] segment generated from a ten state track (solid), and an [ɔʸ] token (+) randomly selected from the test set.

Fixed endpoints with interpolation is a linearly interpolated mapping of a token's frames to the states of the track. The initial and final frames of the token are always aligned with the initial and final states of the track with intermediate frames falling linearly in between. Intermediate frame values are interpolated between adjacent track states. Figure 1 shows Mel-frequency cepstral coefficients C0 thru C3 for a synthetic [ɔʸ] segment and an [ɔʸ] token, randomly selected from the test set.

ERROR MODELLING

The objective of the error model is to take advantage of information residing in the error correlations both over time and between attributes. In [14, 15, 16] the errors are assumed to be independent. This assumption is appropriate for the stochastic dynamic model systems of [14, 15] since the Kalman filter produces a white innovations process. For other methodologies however, the simplicity gained by assuming the error sequence is independent is potentially damaging since the correlation information is discarded. An alternative is to model the error as a single joint-Gaussian distribution [11, 18].

Two key difficulties exist in modelling the error with a Gaussian distribution. The first problem is due to the fact that the error sequence varies in duration. The second problem arises when the dimension of the Gaussian distribution becomes large, and the estimate of the covariance matrix parameters become suspect. One method of overcoming the first problem is by subsampling longer tokens and interpolating short tokens so that all error sequences have a fixed number of frames N. Unfortunately, this can result in a high-dimensional joint-Gaussian if the value of N is large, and the loss of a large amount of data if the value of N is small.

Another method, which we found to provide a good tradeoff, is to allow the error vectors to be of varying frame length and to normalize the frame length by averaging the vectors over each of Q pieces. For example, for a ten state track with Q equal to three, that part of the error which resulted from comparing the token to the first third of the track (i.e. the first three and a third "states") would be averaged, and so on for each of the other two thirds. This technique has the advantages of reducing the dimensionality of the Gaussian pdf while utilizing all of the data. The disadvantage is the loss of the fine detail of the error correlations. However, much of the broader detail

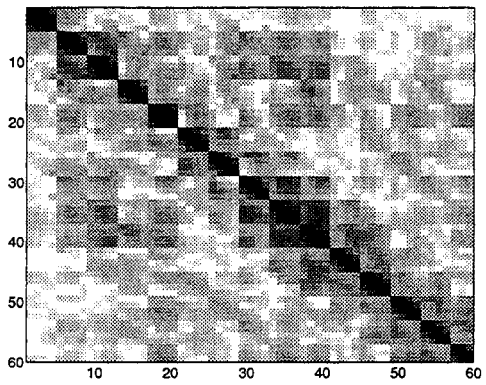

Figure 2: Matrix of error correlation coefficients for the phone [e] with 15 Mel-frequency cepstral coefficients (C0-C14) and $Q = 4$ (4x15 = 60 dimensions). The figure is arranged such that the coefficients of C0 over time are in the upper left, and the C14 coefficients are in the lower right. The absolute value of each element was taken so that large correlations show up dark and areas of little or no correlation show up light. If the errors were assumed independent, the diagonal would have been black and all other elements would be white.

IPA							
ɑ	æ	ʌ	ɔ	ɑʷ	ɑʸ	ɛ	ɝ
e	ɪ	i	o	ɔʸ	ʊ	u	ü
TIMIT							
aa	ae	ah	ao	aw	ay	eh	er
ey	ih	iy	ow	oy	uh	uw	ux

Table 1: Vowels used for the classification experiments.

is retained. Figure 2 shows the absolute values of the error correlation coefficients for the vowel [e] using a value of Q equal to four. Note the high degree of off-diagonal structure, which would be lost under an assumption of independence. It should be noted that normalization of an acoustic vector into a fixed number of parts was also employed in [6, 9].

Therefore, our model of the error is the maximum likelihood estimate of the mean (which is not zero, due to the averaging into Q pieces) and full covariance matrix for each vowel. For P acoustic attributes, the result is a joint-Gaussian density of dimension PQ. It is important to note that the dimension of the model is independent of the number of states used to characterize the track, T.

CLASSIFICATION EXPERIMENTS

Although our ultimate goal is phonetic recognition, we have initially focused on the problem of phonetic *classification*. The experiments reported here are limited to the thirteen monothong and three diphthong vowels of American English, as shown in Table 1. These vowels were chosen for our initial experiments since they can exhibit a large amount of dynamic behavior, and thus provide a good starting point for the evaluation of our methodology. There are also several other published vowel classification studies which can used as a basis of comparison.

Task, Corpus, and Signal Representation

The vowel classification experiments are based on the TIMIT acoustic-phonetic speech corpus [17]. As shown in Table 2, our training and test sets were taken from the phonetically-compact

Type	# Speakers	# Sentences	# Vowel Tokens
Train (*sx*)	499	2,495	20,528
Train (*sx,si*)	499	3,992	34,576
Test	50	250	1,879

Table 2: Training and test sets for the classification experiments.

Condition	Covariance Parameters	Correct (%)
Diagonal	15 x 4 = 60	51.0
Time	15 x (4x4) = 240	55.0
Space	4 x (15x15) = 900	55.8
Full	(4x15) x (4x15) = 3600	62.5

Table 3: Performance as a function of information retained in the covariance matrices.

sx utterances. These were chosen to match those used in previous studies [4, 6, 9, 16]. The effects of adding the *si* utterances to the training data were also investigated.

All of the experiments used Mel-frequency cepstral coefficients (MFCC's) to represent the speech signal [19]. The first fifteen MFCC's were used to facilitate comparison to previously published results [6, 16]. Some of our experiments also explored the use of ΔMFCC's which were computed at the beginning and end of the vowel segment [18].

Experimental Results

The vowel classification results reported here were all based on a common track configuration of ten states. This choice was determined by the results of our distortion studies which showed that the reduction in mean distortion (defined in 4) began to asymptote at larger values. In all cases the resulting error was quantized using a value of four for Q. This was determined by some initial studies on an independent development set which showed consistent improvement as Q is increased from one to four, followed by a gradual degradation for larger values, presumably due to lack of training data. The choice of these parameters resulted in a 60 dimensional distribution for experiments with the fifteen MFCC's, and 90 dimensions for experiments that included the two cepstral differences.

In our first set of experiments we investigated the relative importance of the temporal and spatial correlations in the error for classification. As shown in Table 3 we performed a set of four experiments using 15 MFCC's and a single Gaussian error model for each phone. The *diagonal* condition assumed total independence of all dimensions in the error. The *time* condition retained the temporal correlations of each MFCC across the four sub-segments, and assumed independence between MFCC's. The *space* condition assumed independence between each of the four sub-segments. Thus, it modelled the MFCC correlations within each sub-segment and also captured some temporal information as well, since a separate block was trained for each sub-segment. It is interesting to observe that although the *space* condition utilized nearly four times the number of parameters as the *time* condition, the performance was very similar. This highlights the importance of retaining the temporal correlations. Finally, the *full* condition modelled all correlations, and produced the highest accuracy of 62.5%. This last condition was used for all subsequent experiments.

Our next set of experiments augmented the static MFCC's with additional information. As shown in Table 4 we explored the impact of vowel prior probabilities, duration, and ΔMFCC's

Description		Baseline (%)	Gender (%)
MFCC's		60.9	62.4
+ Priors		62.5	62.5
+ Duration		63.8	65.1
+ ΔMFCC's	Full	65.7	67.1
	Block	66.0	67.5
+ *si* Training Data		66.6	68.9

Table 4: Results for the vowel classification experiments.

on classification performance, and also investigated the effects of adding the *si* utterances to the training set. For each experimental condition we evaluated a *baseline* and a *gender* configuration. The baseline configuration used a single track for each of the sixteen vowels, and a single Gaussian distribution for each error model. The gender configuration augmented the baseline with tracks and error models which were trained separately on male and female speakers. Hence, three models were produced for each vowel. During testing the gender was unknown, and the top-scoring model determined the classification result.

The addition of the ΔMFCC's at the segment boundaries increased the error dimensionality from 60 to 90. Due to this high dimensionality, we explored two alternative techniques to model the error. The full 90 x 90 covariance matrix was used for the *full* experiment. For the *block* experiment we assumed independence between the static 60 dimension MFCC vector and the 30 dimension ΔMFCC vector. This experiment was motivated by our observation that the cross-correlation terms between the MFCC's and the differences were generally small, thus adding little information at a cost of 3600 parameters for each phonetic model. However, when the *si* training data is added, the *block* condition results remain unchanged, but the *full* covariance results improve significantly, and these are the results reported in Table 4.

DISCUSSION

We believe the spectral dynamic models evaluated in this paper show promise for the task of phonetic classification. Our studies have shown that temporal correlations are almost as important for classification as spatial correlations.

The vowel classification results we obtained are competitive with those found in the literature. Meng reports 59.6% on the same task when using fifteen MFCC's with an MLP classifier [20]. Her representation is very similar to our static MFCC experiment which achieved 62.5%, although she did not use C0. Her best result was 65.6% using two auditory model outputs.

Carlson and Glass also reported results on this vowel classification task using an MLP classifier [4]. Their most similar experiment used three average Bark spectral vectors, obtaining 62.5% accuracy. When they included gender information, they obtained 65.8% with a formant-based representation. They found that duration information improved classification performance by around 1.3%, which agrees with our results.

More recently, Gish and Ng have examined this task as part of their evaluation of a segmental speech model. Using MFCC's and their differences, along with the segment duration, they obtained a result of 65.5%. It is interesting to observe that their performance degrades nearly 6% when duration is not used. This result differs with our experience. Without duration, our best gender models (based on *sx* data only) achieve 66.5% classification accuracy (not shown in Table 4).

FUTURE WORK

In the future we hope to more fully exploit the potential of the dynamic tracks by incorporating contextual dependencies. We have started to investigate independently *perturbing* a track to account for both the left and right contexts. Tracks are made for the two contexts separately and then *merged* during generation to create a context-sensitive synthetic segment. Given that we wish to create tracks for L contexts, this requires us to build $2L$ tracks as opposed to L^2 tracks, which could dramatically alleviate the sparse data problem often associated with context-dependent modelling. Additionally, since our error modelling techniques are independent of the track, we need not have $2L$ error models, but can pool the errors over all contexts. We have experimented with this type of independent contextual approach and are able to achieve distortion reductions in the test set in the range of 10 to 20% for a large selection of the vowels with contexts based on place of articulation.

REFERENCES

[1] P. Brown, "The acoustic-modeling problem in automatic speech recognition," Ph.D. Thesis, Carnegie-Mellon University, 1987.

[2] K.F. Lee and H.W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. ASSP*, Vol. 37, No. 11, 1641–1648, 1989.

[3] C.H. Lee, L.R. Rabiner, R. Pieraccini, J.G. Wilpon, "Acoustic modeling for large vocabulary speech recognition," *Computer Speech and Language*, Vol. 4 , 127–165, 1990.

[4] R. Carlson and J. Glass, "Vowel classification based on analysis-by-synthesis," *Proc. ICSLP 92*, 575–578, Banff, Canada 1992.

[5] H.C. Leung, I.L. Hetherington, V.W. Zue, "Speech recognition using stochastic explicit-segment modeling,", *Proc. Eurospeech 91*, 931–934, Genoa, Italy, September, 1991.

[6] H. Meng and V. Zue, "A comparative study of acoustic representations of speech for vowel classification using multi-layer perceptrons," *Proc. ICSLP 90*, 1053–1056, Kobe, Japan, November, 1990.

[7] M. Phillips, and V. Zue, "Automatic discovery of acoustic measurements for phonetic classification," *Proc. ICSLP 92*, 795–798, Banff, Canada, October, 1992.

[8] H.C. Leung, B. Chigier, J.R. Glass., "A comparative study of signal representations and classification techniques for speech recognition," *Proc. ICASSP 93*, 680–683, Minneapolis, MN, April, 1993.

[9] H. Leung and V. Zue, "Phonetic classification using multi-layer perceptrons," *Proc. ICASSP 90*, pp. 525–528, Albuquerque, NM, April, 1990.

[10] V. Zue, J. Glass, M. Phillips, S. Seneff, "Acoustic segmentation and phonetic classification in the SUMMIT system," *Proc. ICASSP 89*, 389–392, Glasgow, Scotland, 1989.

[11] M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition," *IEEE Trans. ASSP*, Vol. 4, No. 12, 1857–1869, December, 1989.

[12] S.E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language*, Vol. 1, No. 1, 24–46, March, 1986.

[13] M. Russel, "A segmental HMM for speech pattern modelling," *Proc. ICASSP 93*, 499–502, Minneapolis, MN, April, 1993.

[14] V. Digilakis, J. Rohlicek, M. Ostendorf, "A dynamical system approach to continuous speech recognition," *Proc. ICASSP 91*, 289–292, Toronto, Canada, May, 1991.

[15] V. Digilakis, "Segment-based stochastic models of spectral dynamics for continuous speech recognition." Ph.D. Thesis, Boston University, 1992.

[16] H. Gish and K. Ng, "A segmental speech model with applications to word spotting," *Proc. ICASSP 93*, 447–450, Minneapolis, MN, April, 1993.

[17] L. Lamel, R. Kassel, S. Seneff, "Speech database development: design and analysis of the acoustic-phonetic corpus," *Proc. DARPA Speech Recognition Workshop*, Report No. SAIC-86/1546, 100–109, February, 1986.

[18] B. Chigier and H. Leung, "The effects of signal representations, phonetic classification techniques, and the telephone network," *Proc. ICSLP 92*, 97–100, Banff, Canada, October, 1992.

[19] P. Mermelstein and S. Davis, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. ASSP*, Vol. 23, No. 1, 67–72, February, 1975.

[20] H. Meng, "The use of distinctive features for automatic speech recognition," S.M. Thesis, Massachusetts Institute of Technology, September 1991.