First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Detection and Classification of Phonemes Using Context-Independent Error Back-Propagation[1]

Hong C. Leung, James R. Glass
Michael S. Phillips, and Victor W. Zue
Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139, U.S.A.

ABSTRACT

Over the past few years, we have been investigating the problem of utilizing artificial neural networks for phonetic classification. In this paper, we will describe several extensions to our earlier work, utilizing a segment-based approach. We will formulate our segmental framework and report our study on the use of multi-layer perceptrons for detection and classification of phonemes. Issues related to computational requirements and input representations will also be discussed. Our investigation is performed within a set of experiments that attempts to recognize 38 vowels and consonants in American English independent of speaker. When evaluated on the TIMIT database, our system achieves an accuracy of 56%.

INTRODUCTION

Recently, we have been investigating the use of artificial neural networks (ANN) for phonetic *classification*. That is, given a time region in an utterance, the network is asked to identify the phonetic unit in it. Our study was performed on the constrained task of using multi-layer perceptrons (MLP) to classify different speech sounds in American English [1,2]. When evaluated on 38 vowels and consonants, our network achieved a classification accuracy of about 70% [1].

Thus far, the neural networks research community has placed heavy emphasis on the problem of pattern classification. In many applications, including speech recognition, one must also address the issue of *detection*. Thus, for example, one must detect the presence of phonetic segments as well as classify them. Recently, the community has moved more towards recognition of continuous speech. A network is typically used to label every frame of speech in a frame-based recognition system [3,4,5].

Our goal is to study and exploit the capability of ANN for speech recognition, based on the premise that ANN may offer a flexible framework for us to utilize our improved, albeit incomplete, speech knowledge. As an intermediate milestone, this paper extends our earlier work on phonetic classification to context-independent phonetic recognition. Thus we need to locate as well as identify the phonetic units. Our system differs from the majority of approaches in that a segmental framework is adopted. The network is used in conjunction with acoustic segmentation procedures to provide a phonetic string for the entire utterance.

The organization of this paper is as follows. In the next section, we will formulate our segmental framework and address some of the computational issues. We will then describe our experiments on phonetic classification and recognition. Finally, we will report some results and discuss some of our recent studies on using different input representations.

SEGMENTAL FORMULATION

In our segmental framework, a phonetic unit is mapped to a segment explicitly delineated by a begin and end time in the speech signal. Segmental frameworks have been investigated by others [6,7,8] and contrast with the prevailing frame-based structure used by most HMM systems [9], where sequences of observation frames are assumed to be statistically independent from each other. We believe that a segmental framework offers us more flexibility than is afforded by a frame-based approach, and could ultimately lead to superior modelling of the temporal variations in the realization of underlying phonological units. It is for this reason that we base our system on such an approach.

Let $\hat{\alpha}$ denote the best sequence of phonetic units in an utterance. We have taken a stochastic segment approach so that the probability of the best sequence, $p(\hat{\alpha})$, is maximized. Specifically,

$$\hat{\alpha} = \arg\max_{\vec{s}} \prod_{s_i \epsilon \vec{s}} p(\alpha_j|s_i)p(s_i); \qquad 1 \leq j \leq N_\alpha \qquad (1)$$

where $\vec{s}$ is any possible sequence of time segments consisting of $\{s_1, s_2, ...\}$, $p(\alpha_j|s_i)$ is the probability of observing a phoneme, α_j, in a time segment, s_i, $p(s_i)$ is the probability of a valid time segment, and N_α is the number of possible phonetic units. In order to perform recognition, the two probability measures in Equation 1 must be estimated. The first term, $p(\alpha_j|s_i)$, is a set of a-posteriori probabilities and thus can be viewed as a classification problem. The second term, $p(s_i)$, is a set of probabilities of valid time regions and thus can be estimated as a segmentation problem.

Segmentation

In order to estimate the segment probabilities, $p(s_i)$, in Equation 1, we have formulated segmentation into a boundary classification problem. Let $\{b_1, b_2, .., b_K\}$ be the set of boundaries that might exist within a time segment, s_i, as shown in Figure 1a. These boundaries can be proposed by a boundary detector, or they can simply occur at every frame of speech. We define $p(s_i)$ to be the probability that all the boundaries within s_i do not

[1] This research was supported by DARPA under Contract N00014-89-J-1332, monitored through the Office of Naval Research.

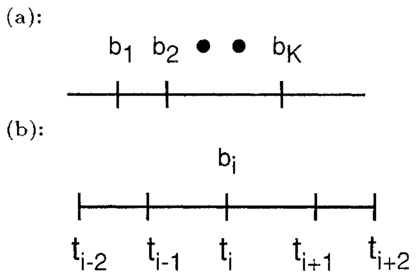

Figure 1: Schematic diagrams for estimation of (a) segment probability, $p(s_i)$, and (b) boundary probability, $p(b_k)$. The boundaries can be proposed by a boundary detector, or they can simply occur at every frame. See text.

exist. To reduce the complexity of the problem, assume b_k is statistically independent of b_l for $\forall k \neq l$. Thus,

$$\begin{aligned} p(s_i) &= p(\bar{b}_1, \bar{b}_2, .., \bar{b}_K) \\ &= p(\bar{b}_1)p(\bar{b}_2)...p(\bar{b}_K) \end{aligned} \qquad (2)$$

where $p(\bar{b}_k)$ stands for the probability that the k^{th} boundary does not exist. As a result, the probability of a segment, $p(s_i)$ can be obtained by computing the probabilities of the boundaries, $p(b_k)$, within the segment. As we will discuss in a later section, by using the time-aligned transcription, we can train the boundary probabilities in a supervised manner.

Phonetic Classification

Once the probability of a segment, $p(s_i)$, is obtained, we still need to classify it, i.e. compute the probabilities of the phonetic units in the segment, $p(\alpha_j \mid s_i)$. Again, the time-aligned transcription can be used to train the probabilities in a supervised manner. We have discussed this in earlier papers [1,2]. In a later section, we will discuss some of our recent experimental results.

Computational Requirements

One of the disadvantages of our segmental framework is that the amount of computation involved can be very significant. Let N denote the number of frames in an utterance. The number of possible segments, N_s, is, therefore, $N(N-1)/2$, resulting in 2^{N-2} distinct segmentations. Since phonetic classification is needed in each possible segment, the amount of computation and the size of the search space can be prohibitively large. An example is illustrated in Table 1. If an utterance is 3 seconds long, and is analyzed once every 5 msec., $N_s \approx 180,000$.

Several techniques can be adopted to reduce the computational requirement. First, it is clear that certain search techniques, such as those that involve Viterbi search, can reduce the amount of computation. In the following paragraphs, we discuss some other techniques that we have explored.

Boundary Pruning As we have previously discussed, a set of boundaries can first be proposed by a boundary detector. Let N_b denote the number of boundaries proposed by a boundary detector in an utterance. If a segment is constrained to have its end points located at the proposed boundaries, the number of possible segments, N_s, is equal to $N_b(N_b-1)/2$. If N_b is small compared to N, N_s can be significantly reduced. In the SUMMIT system for instance [8], $N \approx 5N_b$, thus reducing N_s by more than an order of magnitude. If an utterance is 3 seconds long, $N_s \approx 7,200$.

	N_s	Example
Exhaustive Enumeration	$\frac{N(N-1)}{2}$	180,000
Boundary Pruning	$\frac{N_b(N_b-1)}{2}$	7,200
Binary Hierarchy	$2N_b-3$	240

Table 1: Computational requirements for segment-based frameworks. N stands for the number of frames in an utterance, N_b stands for the number of boundaries proposed in the same utterance. Example is based on an utterance of 3 seconds long, $N = 5N_b$, and an analysis rate of 200 times/second.

Segment Pruning There are many alternatives that can be used to reduce the number of segments. For example, Kopec and Bush used conservative duration estimates to eliminate many candidate segments [6]. In the SUMMIT system, a binary hierarchy called dendrogram is constructed based on some proposed boundaries. Regardless of the specific implementation, such a binary hierarchical representation results in $N_s = 2N_b - 3$. Again, if an utterance is 3 seconds long, and $N = 5N_b$, then $N_s \approx 240$, a reduction in computation by almost 3 orders of magnitude.

Thus there is a continuum in the computational requirement for a segment-based recognition system. By adopting different pruning techniques, the amount of computation can be made more manageable. However, the performance of the overall recognition system depends on the reliability of the pruning techniques and the robustness of the boundary detector. In the next section, we will describe experiments and compare results using different techniques.

EXPERIMENTS

In the previous section we outlined our segmental formulation and contrasted the computational requirements needed for different techniques. In this section we will discuss some experiments using MLP. First, we will describe our task and corpus. We will also review our work on phonetic classification and discuss how acoustic segmentation can be performed using MLP. Finally, we will report our phonetic recognition results.

Tasks and Corpora

The experiments described in this paper deal with classification and recognition of 38 phonetic labels representing 14 vowels, 3 semivowels, 3 nasals, 8 fricatives, 2 affricates, 6 stops, 1 flap and 1 silence. This particular set was chosen because it has been used in other recent evaluations within and outside our research group. Within the context of classification, the networks are given a segment of the speech signal, and are asked to determine its phonetic identity. Within the context of recognition, the networks

Corpus	Set	Speakers	Sentences	Tokens	Type
I	training	300	1500	55,000	sx
	testing	50	250	9,000	sx
II	training	500	4000	150,000	sx/si
	testing	50	400	15,000	sx/si

Table 2: Corpora I and II extracted from the TIMIT database. Corpus I contains only sx sentences, whereas Corpus II contains both sx and si sentences. The speakers in the testing sets for both Corpus I and Corpus II are the same.

are given an utterance, and are asked to determine the identity and locations of the phonetic units in the utterance. All experiments were based on the sentences in the TIMIT database [10]. As summarized in Table 2, Corpus I contains 1,750 sx sentences spoken by 350 male and female speakers, resulting in a total of 64,000 phonetic tokens. Corpus II contains 4,400 sx and si sentences spoken by 550 male and female speakers, resulting in a total of 165,000 phonetic tokens.

Phonetic Classification

As previously discussed, estimation of the a-posteriori probability, $p(\alpha_j \mid s_i)$ in Equation 1 can be viewed as a classification problem. Many statistical classifiers can be used. We have chosen to use the MLP, due to its discriminatory capability, as well as its flexibility in that it does not make assumptions about specific statistical distributions or distance metrics. In addition, earlier work by Bourlard and Welleken shows that the outputs of MLP can approximate a-posteriori probabilities [11].

In classifying the 38 vowels and consonants, we discovered some major problems in training the network. These problems were subsequently overcome by procedures such as judicious initialization to enable the network to learn quickly and converge to a better local minimum, normalization of inputs to enhance learnability of the network, adaptive gain to enable the network to pay similar attention to different phonetic units, and modular training to reduce training time [1].

There were two representations used as input for the MLP classifier. The first representation was identical to that in the SUMMIT system, and consisted of 82 acoustic attributes. These segmental attributes were generated automatically by a search procedure that uses the training data to determine the settings of the free parameters of a set of generic property detectors using an optimization procedure[12]. The second representation consisted of a vector of three average spectra which corresponded to the left, middle, and right thirds of a segment. The spectra were the mean-rate and synchrony outputs of a 40 channel auditory model [13]. Thus, there were 120 dimensions used for each representation. Finally, segment duration was also included.

Boundary Classification

In our segmental framework formulated in Equation 1, the main difference between classification and recognition is the incorporation of a probability for each segment, $p(s_i)$. As described previously in Equation 2, we have simplified the problem of estimating $p(s_i)$ to one of determining the probability that a boundary exists, $p(b_k)$.

To estimate $p(b_k)$, a MLP with two output units is used, one for the valid boundaries and the other for the extraneous boundaries. By referencing the time-aligned phonetic transcription, the desired outputs of the network can be determined. In our current implementation $p(b_k)$ is determined using four abutting segments, as shown in Figure 1b. These segments are proposed by the boundary detector in the SUMMIT system. Let t_i stand for the time at which b_i is located, and s_i stand for the segment between t_i and t_{i+1}, where $t_{i+1} > t_i$. The boundary probability, $p(b_i)$, is then determined by using the average mean-rate response in s_{i-2}, s_{i-1}, s_i, and s_{i+1} as inputs to the MLP. Thus the network has altogether 160 input units.

Results

Phonetic Classification In the phonetic classification experiments, the system classified a token extracted from a phonetic transcription that had been aligned with the speech waveform. Since there was no detection involved in these experiments only substitution errors were possible.

In the first set of experiments, we compared results based on Corpus I, using different classifiers and different input represenations. As has been reported previously, the baseline speaker-independent classification performance of SUMMIT on the testing data was 70% [8]. We also experimented with representations based on the spectral outputs described previously. Four experiments were performed using 1) the synchrony outputs, 2) the mean-rate outputs, 3) the synchrony and mean-rate outputs and 4) the synchrony and mean-rate outputs and segment duration. The results of all experiments have been summarized in Table 3. Finally, we used the same set of acoustic attributes used in the SUMMIT system. The MLP classifier yields a performance of 74%.

These results collectively suggest that the MLP classifier produces results favorable to those in the current SUMMIT system. Furthermore, the use of the automatically determined acoustic attributes can lead to a better classification performance.

Although the sx sentences were designed to be phonetically balanced, the 1,750 sentences in Corpus I are not distinct. In the second set of experiments, we evaluated the MLP classifier on Corpus II, which include both the sx and si sentences.[2] As shown in Table 3, the classifier achieves 76%.

Connections All the networks used as described in Table 3 have only 1 hidden layer. They have a different number of input units, depending on the acoustic representations. For example, when both the synchrony envelopes and mean-rate response are used, the network has a total of 240 input units. The networks may also have a different number of hidden units, resulting in a different number of connections. As we can see from Figure 3, the attributes have the additional advantage over the raw auditory outputs in that the number of connections needed is decreased. In other words, training time, classification time, and storage requirements can all be reduced.

Boundary Classification We have evaluated the boundary classifier using the training and testing data in Corpus I. By using 32 hidden units, the network can classify 87% of the boundaries in the test set correctly.

[2] All the si sentences in TIMIT are distinct.

	Classifier	Representation	Correct	Connections
I	SUMMIT	attributes	70%	-
I	MLP	SYN	65%	10,000
I	MLP	MR	68%	10,000
I	MLP	SYN+MR	70%	35,000
I	MLP	SYN+MR+DUR	72%	40,000
I	MLP	attributes	74%	15,000
II	MLP	attributes	76%	30,000

Table 3: Phonetic classification comparing the baseline and MLP classifers, and acoustic representations. The representations are the 82 acoustic attributes, the synchrony envelopes (SYN), mean-rate response (MR), and duration (DUR). Also shown are the number of connections in the networks.

Corpus	Classifer	Segment	Correct
I	Baseline	Binary Hierarchy	47%
I	MLP	Binary Hierarchy	50%
I	MLP	Boundary Pruning	55%
II	MLP	Boundary Pruning	56%

Table 4: Phonetic recognition results using binary hierarchy (dendrogram), and boundary pruning. No duration, bigram, or trigram statistics have been used. Errors include substitutions, deletions, and insertions.

Phonetic Recognition The results of the phonetic recognition experiments are shown in Table 4. No duration, bigram, or trigram statistics have been used. The baseline performance of the current SUMMIT system on Corpus I is 47%, including substitution, deletion, and insertion errors. When the MLP was used in place of the classifier in the current SUMMIT system, the performance improved to 50%. When the MLP was used with the boundary pruning technique discussed in the previous section, the segments were further pruned based on conservative duration constraints. As a result, they contained about twice as many regions, on average, as the binary hierarchy. As can be seen from Table 4, the performance improved to 55%. Finally, by using the network trained and tested on Corpus II, the performance improved to 56%.

DISCUSSION

Equation 2 shows one way of estimating the segment probability, $p(s_i)$. More recently, we have been investigating other procedures. One alternative is to utilize both the internal and external boundaries. Thus Equation 2 can be modified to:

$$\begin{aligned} p(s_i) &= p(b_l, \bar{b}_1, \bar{b}_2, .., \bar{b}_K, b_r) \\ &= p(b_l)p(\bar{b}_1)p(\bar{b}_2)...p(\bar{b}_K)p(b_r) \end{aligned} \qquad (3)$$

where b_l and b_r stand for the boundaries at the left and right end points of a segment, s_i, as shown in Figure 2. When $p(b_l)$ and $p(b_r)$ approach 1, Equation 2 approximates Equation 3. Conceptually, Equation 3 should result in a better estimate of $p(s_i)$. We have been conducting experiments in this direction, and are hopeful that we can report more results in the future.

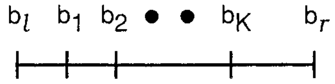

Figure 2: Schematic diagram for alternative way of estimating segment probability, $p(s_i)$. See text.

In summary, we have discussed issues in adopting a segmental framework. Although it offers us more flexibility than is afforded by a frame-based approach, its computational requirements can be prohibitively large. We have discussed several pruning techniques, and have shown that there is a continuum in computational requirement. We have also discussed using the MLP to perform context-independent phonetic classification and detection, using different input representations. We have shown that the MLP yields results favorable to the classifier in the current SUMMIT system, and that the use of acoustic attributes results in improved performance and reduced computations. Future work includes the use of context-dependent models for phonetic and boundary classification, utilization of other phonological units, and extension to recognition of continuous speech.

REFERENCES

[1] Leung, H.C., and V.W. Zue, "Phonetic Classification Using Multi-Layer Perceptrons," *Proc. ICASSP-90*, Albuquerque, 1990.

[2] Leung, H.C., *The Use of Artificial Neural Networks of Phonetic Recognition*, Ph.D. Thesis, Mass. Inst. of Tech., 1989.

[3] Franzini, M.A., K.F. Lee, and A. Waibel, "Connectionist Viterbi Training: A New Hybrid Method for Continuous Speech Recognition," *Proc. ICASSP-90*, Albuquerque, NM, USA, 1990.

[4] Morgan, N., and H. Bourlard, "Continuous Speech Recognition Using Multilayer Perceptrons with Hidden Markov Models," *Proc. ICASSP-90*, Albuquerque, NM, USA, 1990.

[5] Tebelskis, J., and A. Waibel, "Large Vocabulary Recognition Using Linked Predictive Neural Networks," Proc. ICASSP-90, Albuquerque, NM, USA, 1990.

[6] Kopec, G.E., and M.A. Bush, "Network-Based Isolated Digit Recognition Using Vector Quantization," *IEEE Trans. Acoust. Speech and Sig. Proc.*, Vol. ASSP-23, pp. 850-867, 1985.

[7] Ostendorf, M., and S. Roucos, "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Trans. Acoust. Speech and Sig. Proc.*, Vol. 37, No.12, pp. 1857-1869, 1989.

[8] Zue, V., J. Glass, M. Phillips, and S. Seneff, "The MIT SUMMIT Speech Recognition System: A Progress Report," *Proceedings of DARPA Speech and Natural Language Workshop*, February, 1989.

[9] Lee, K.F., *Automatic Speech Recognition: The Development of the SPHYNX System*, Kluwer Academic Publishers, Boston 1989.

[10] Lamel, L.F., R.H. Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic Phonetic Corpus," *Proc. DARPA Speech Recognition Workshop*, 1986.

[11] Bourlard, H., and C.J. Wellekens, "Links between Markov Models and Multilayer Perceptrons," *Advances in Neural Information Processing Systems 1*, Morgan Kaufmann.

[12] Phillips, M.S., "Automatic Discovery of Acoustic Measurements for Acoustic Classification," *J. Acoust. Soc. Amer.*, Vol. 84, 1988.

[13] Seneff, S. "A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing," *Proc. J. of Phonetics*, 1988.