Connectionist Large Vocabulary Word Recognition

A. Waibel

26 October 1989

Carnegie-Mellon University
ATR Interpreting Telephony Research Laboratories

Table of Contents

Abstract

In this paper we discuss recent research aimed at extending connectionist models to large vocabulary word recognition. We describe the problem and the properties a successful large vocabulary system must satisfy. While a number of different methods and ideas have recently been proposed and are under investigation we will limit the discussion here to only one particular hybrid approach, i.e., the combination of TDNN-based phoneme recognition/spotting nets with classical techniques for sequence management (such as DP-matching and HMMs). We implement a baseline system using the best recent TDNN phoneme spotting nets and evaluate its performance over a 500 and a 2620 word vocabulary, not used during training. In both of these vocabulary independent evaluations high word recognition rates were measured despite the large vocabulary size and perplexity in this otherwise unconstrained task. We then describe, exploratory experiments that illustrate the importance and effectiveness of integral training, i.e., the integration of sequential management or alignment with phoneme network optimization. Significant performance improvements were found with this technique over a system using decoupled training and alignement. Finally, we offer a critical discussion and observations for further research.

Acknowledgement

The author gratefully acknowledges the help and support of Dr. Akira Kurematsu, Dr. Kiyohiro Shikano and the ATR Interpreting Telephony Research Laboratories, that made this research possible.

1 Introduction

Connectionist models have recently attracted considerable attention as approach for the design of large vocabulary speech recognition. Underlying this interest is the perception that connectionist models or "neural networks" are not only attractive for efficient hardware implementations, but that they also deliver very high recognition performance, frequently exceeding the performance of traditional speech recognition techniques [Waibel 87, Waibel 88a, McDermott 89, Waibel 88b]. These performance results were achieved by the ability of these networks to optimally adjust the interconnections between massively parallel and distributed simple processing elements, thus mimicking some of the processing properties of neural processing. Although earlier work [Elman 87] had already highlighted interesting abstractions that such networks learned in order to perform their tasks, further advances were required to deal with the temporal nature of speech: the dynamic properties of the speech signal and the need for segmentation free processing. A variety of techniques were reported [Tank 87, Waibel 87, Watrous 87]. The Time-Delay Neural Network proposed solutions to these requirements by introducing time-delayed connections and shift-invariant learning and recognition [Waibel 87, Waibel 89a]. Excellent performance was achieved with these networks even at the absence of segmentation, i.e., even when such networks were applied to phoneme spotting in running speech [Sawai 88, Sawai 89]. Most recent results indicate that phoneme spotting rates of up to 98% can be achieved for 23 Japanese phonemes over a Japanese 5240 word large vocabulary database [Sawai 90]. Similarly, it has been shown [Bottou 88, Bottou 89] that segmentation free *word* recognition at high recognition performance can be achieved with similarly structured time-delay neural networks. These networks however, posess output units corresponding to each word in the recognition vocabulary and required training examples for each of the vocabulary words. Clearly, for large vocabulary recognition, word models based on *subword units* must be developed. This is one of the topics addressed in this report. The second emerging question discussed in the following is the question of robustness: It is desirable for networks to both recognize speech at high recognition rates as well as having them degrade gracefully, under potential changes in recording conditions, noise, acoustic transducer, speaking rate and style, speaker and task requirements.

2 Models for Large Vocabulary Word Recognition

The recognition of words in large vocabularies is complicated by two distinct problems: First, large vocabularies tend to become more confusable as more and more words crowd the acoustic space and become acoustically increasingly similar. Second, it is impractical to train large vocabulary systems, one word at a time, due to the enormous amount of training data and computation that such an approach would require, not to mention the difficulty of adding new vocabulary items. Clearly, successful practical systems must be based on *subword units*, such as phonemes, diphones or syllables. These constraints have motivated a large body of research aimed at high performance phoneme recognition. Statistical solutions (HMMs) have been proposed, that have the advantage of being easily combined into words and sentences, but have been limited in their ability to handle fine acoustic discriminatory detail. Neural networks solutions on the other hand have produced excellent recognition performance at the acoustic phonetic level, but only preliminary attempts have been made so far in integrating these phoneme models into words and sentences. This is due in part to the still exploratory nature of models aimed at connectionist sequential processing in general (see for example [Wong 86, Elman 88, Servan-Schreiber 88]) that are only beginning to mature into large performance system implementations. Most current activity in connectionist speech recognition is therefore aimed either at combining connectionist and classical methods, or at the development of novel connectionist extensions towards integrated pattern sequence processing. In the following we describe some of these techniques and report in detail on results from initial large vocabulary isolated word recognition experiments.

3 Hybrids: Neural Networks and Classical Techniques

Most popular at present perhaps are so-called hybrid models, that seek to combine the perceived strengths of connectionist models with those of more classical recognition techniques such as Dynamic Programming or Hidden Markov Models. Under this approach, connectionist models are viewed as high performance non-linear classifiers that could replace more rudimentary distance metrics, or vector quantization steps commonly found at the front end of most typical recognizers. Dynamic Programming algorithms and Hidden Markov Models are then viewed as mechanisms to provide sequence management, i.e., impose the additional constraints that phonemes must occur in the right order to be producing a legal word. This paper explores some models of this kind and we describe initial results in

the following.

In the approach explored here, a TDNN is chosen to classify input speech into one of several possible phoneme output categories. The experiments reported here are based on a Japanese large vocabulary isolated word database described elsewhere [Sagisaka 87, Waibel 87, Waibel 89a]. As before phonemes from this database were used to train TDNNs to produce one of 24 phoneme output categories (5 vowels, 18 consonants and silence) as speech flows by. Due to the shift-invariance property of TDNNs these networks have also been demonstrated to produce high performance for phoneme spotting [Waibel 89b], recently 98% in speaker-dependent open test phoneme spotting experiments [Sawai 90].

3.1 From Phonemes to Words

The most straight forward approach to integrating such connectionist networks into large vocabulary recognition systems is depicted in Fig.1. Here the output of a set of 24 TDNNs is used in form of a vector

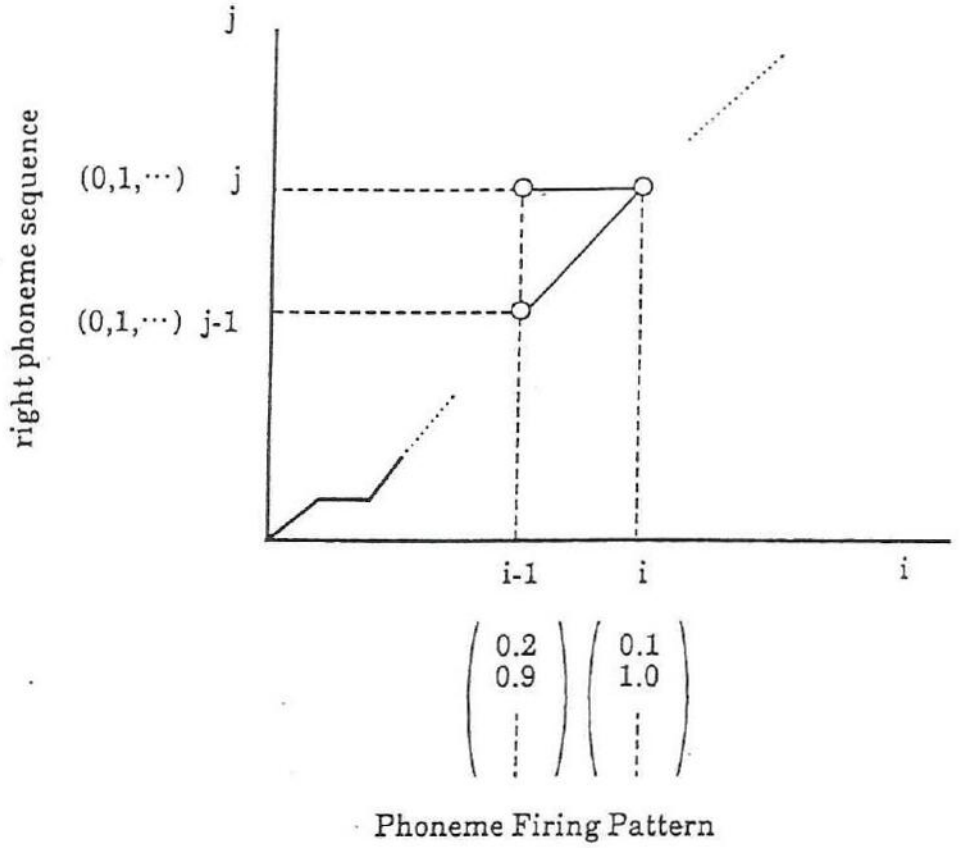

Figure 1: DP-Based Sequential Management

sequence of phoneme hypotheses every 20 msec. In the experiments reported here, these vectors are

obtained by averaging two consecutive TDNN output activations sampled every 10 msecs. DP-matching [Sakoe 78, Itakura 75] is then applied to align these phoneme outputs with a target vector representing one of the candidate words' dictionary phoneme sequence. A Euclidean distance is computed along the best alignment path and the sum of all local distances on this path constitutes the word score for the word candidate. Naturally, this simple approach is limited in a number of ways: First, Dynamic Programming matching ultimately may not be the optimal method one might like to choose for sequence management. It assumes that different phonemes are independent states and do not interact with each other[1]. Second, there is no guarantee that a Euclidean distance is the most useful measure here given that output activations really simulate binary classification decisions. Perhaps the most important limitation is that word alignment, word score computation and training of the underlying TDNNs had been performed independently. Hence there is no guarantee, that what was a priori defined as a speech frame for phoneme X, is indeed the best assignment in view of optimal word recognition. Before addressing some of these limitations in the section below, however, we first report results from benchmarking experiments as a baseline for further work. It should be noted, however, that the simplicity of this approach does have two advantages: First, it is phoneme based, and hence extendable to large vocabulary recognition. Second, phonemes are trained over a large training database and the trained TDNNs incorporate a large amount of typical acoustical variations. This is advantageous in view of vocabulary indepdendent recognition, i.e., recognition of words that have not been considered during the training phase, a problem that remains largely unsolved in speech recognition technology to date.

3.2 Vocabulary Independent Recognition Experiments

A testing vocabulary different from the training data and training vocabulary was used for these experiments. A set of rules converted the romaji spelling of -initially- 500 and then 2620 test words into pseudo-phonemic transcriptions. Test speech utterances for each of these test words were then run through our phoneme spotting TDNNs. The outputs of these nets were then aligned with the phoneme sequence of each test word and the best matching word selected as recognition result. Table 1 shows the results from the 500 word and 2620 word recognition experiments, respectively.

---

[1]The TDNNs alleviate this problem somewhat by the fact that adjacent output activations are obtained in part by inclusion of overlapping speech input information.

1st choice	2nd choice	3rd choice	4th choice	5th choice	test vocabulary
96.8%	99.2%	99.8%	99.8%	99.8%	500 words
97.4%	99.2%	99.8%	99.8%	99.8%	500 words (homophones eliminated)
90.4%	95.6%	97.1%	97.7%	98.2%	2620 words
93.9%	96.4%	97.5%	98.0%	98.4%	2620 words (homophones eliminated)

Table 1: Preliminary Baseline Hybrid DP-TDNN Word Recognition Results

3.3 Error Analysis

Error analysis of these results revealed that a large number of misrecognitions are simply caused by the presence of a fairly large number of homophones in the database[2]. These include mutiple entries for words such as "kizuku", "kata", "seki", with identical phonetic spelling and presumably identical pronounciation. They also include a large number of words that might be distinguishable by prosodic cues, but cannot be identified on the basis of phonetic information alone. Examples are: "ho" vs. "hoo" or "hoshi" vs. "hooshi" (duration) and "hashi" vs. "hashi" (different accentuation). Better duration control or analysis of the accent patterns might provide the means to possibly capture even these distinctions. Since we are limiting ourselves to phoneme based recognition here, we also report in table 1 the performance results that are obtained for the same task, when homophone confusions are eliminated. As can be seen from table 1, these homophone confusions are a most noticable cause for near miss confusions (first vs. second or third choice rates).

Duration control might indeed also eliminate a number of errors that are still part of the performance results reported here. Errors such as "ashi" vs. "okashii" arise presumably from excessive time warping, allowing the recognition procedure to skip over an entire syllable with little penalty. Recognition should be constrained by better models of what constitutes the reasonnable or likely duration of each phoneme.

Further error analysis, finally, shows that many errors are also caused by phonemic misspelling of the

[2]Note, that the database consists of recorded utterances from a *dictionary* of 5240 most common Japanese words.

target words or impropper loading of target phoneme sequences into the dp-alignment vectors. Among them are confusions such as "nyuuse" vs. "musuu", for example, caused by "nyuusu" being transcribed erroneously as "nuusu". These and other errors should of course be eliminated by debugging the lexical representations. Some remaining errors might potentially also be eliminated by the introduction of alternate phonemic transcriptions to represent possible alternate pronounciations for some words. No attempt was made here to control for these errors, however, and they are still part of the errors measured in table 1.

3.4 Discussion

As the foregoing discussion shows, good performance was achieved for a large vocabulary and large perplexity task with even this decoupled strategy that treats alignment/sequence-management and phonemic classifications as distinct and decoupled processes. Nonetheless, as research in speech recognition has shown time and again, such decoupling generally leads to poorer and less robust performance as the definition of a phoneme and the objective of a classifier trained to detecting it may not be optimal in view of the global goal, word recognition[3]. Initial experiments with integral training have already been reported for small (digits) vocabulary tasks [Sakoe 89] We have therefore begun to extend these basic hybrid large vocabulary models towards fully integrated training, more specifically, towards word level optimization of subword units.

4 Integral Training

One of the central ideas in the TDNN is the integration of increasingly abstract sets of features into an output decision, independent of *where* in the input speech these features actually occurred (hence shift-invariance). This position independence is achieved by integrating position dependent local phoneme decisions over a certain input range (in the original TDNN, 150 msec) and passing the combined activation through a sigmoid function. This output nonlinearity deemphasizes local position dependent perturbations in classification performance and only focusses on salient, important features, *anywhere* in the input range. Now, if the actual input range contains not simply one phoneme to be recognized, but a

[3]Although no system currently exists that does this successfully, the same argument actually applies beyond the word level. Ultimately, optimal transmission of ideas is the goal of speech communication, not syntax, words or phonemes.

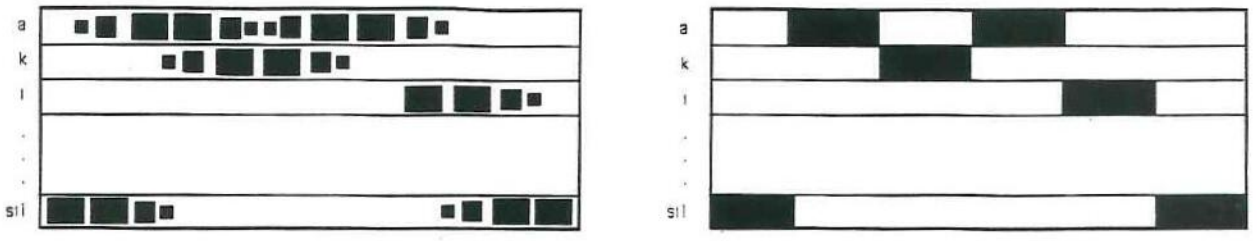

Figure 2: TDNN-target-mask

sequence of phonemes (i.e., a word), then integration has to be performed over varying ranges and compared with changing targets. Fig.2 illustrates the basic mechanism. A set of output activations (shown in the left of Fig.2 are compared with a target mask (shown on the right) and deviations from these targets are corrected by backpropagating error into the underlying TDNNs that have produced the corresponding outputs at that point. Of course, the target mask has to be determined before these assignements and error corrections can take place. Naturally, they could be determined using a database of handlabels, but this would not lead to global optimality and has the unattractive property that precise labels have to exist for all training tokens. Rather, we let DP-alignement produce a set of *hypothesized* boundaries *dynamically*, as training progresses. Error backpropagation is then performed based on these boundaries and the associated target-mask. Using the emerging new set of weights then DP-alignement is performed again, and the process iterates. In this fashion, alignment can seek out optimal transition points as training iterates over many instances of phonemes in different contexts. This approach has so far been partially implemented, and feasibility could be partially demonstrated as

described in the following.

4.1 Method

As before a set of phoneme spotting TDNN's was used as an initial set of phoneme models. The set of phoneme spotting nets used in the experiments here were unfortunately an earlier version of the nets described in the previous section. They had not yet been developed to the level of performance described above. The absolute values of the initial results reported below are therefore somewhat lower than before but the *relative* performance gains achieved based on these networks using integral word level training are nevertheless insightful.

We should also note that several experimental limitations have been imposed to reduce training time and set-up time during this set of exploratory experiments. First, our experiments here were limited to a 225 word vocabulary made up of words that make use of only 10 frequently occurring phonemes, namely the vowels /a,i,u,e,o/ and the consonants /t,k,h,r,s/. The resulting vocabulary, can then be expressed by different combinations of these phonemes alone. It is also likely to lead to more acoustic confusability, which might contribute to lower recognition results, but again, it is the *relative* performance that we are trying to asess here. To further limit the training amount in these experiments, we are also not considering backpropagating errors all the way to the very signal level. Rather, we are taking a modular approach [Waibel 89c], by keeping lower level connections fixed and only apply a higher level net which is dynamically adjusted during word level training and uses the output firings of phoneme spotting nets as input.

4.2 Recognition Experiments

Several experiments were carried out and are described here. Further in depth study of additional experiments can be found in Hirai [Hirai 89].

- We begin by simply aligning the outputs from phoneme spotting nets with word target phoneme sequences. This basically is the approach described in the previous section but evaluated using the earlier (lower performance) nets and the limited vocabulary as described to provide a basis for comparison for the following.
- We then apply a simple higher level network, consisting of one single layer of units without time-delays that connect phoneme spotting outputs to yet another layer of phoneme spotting

units. These units are trained statically on the desired output phoneme targets.

- Third, another simple higher level network is applied, and again trained in a static (non-integral) fashion to test the ability of a network to smooth the underlying phoneme spotting results for word recognition. In contast to the previous higher level net, here time-delays where used to allow the higher level net to smooth the phoneme spotting tracks of the underlying TDNNs over 50 msecs worth of time.

- Finally, dynamic alignment training was introduced. Time-delay (50 msecs) higher level units are first trained statically, and then dynamically for an additional period of 1000 iterations. During these 1000 iterations alignement was done repetitively to change the boundaries of the target mask depicted in Fig.2 with changing weights on the incoming networks[4]. Care was taken that the total (static and dynamic) training was performed over the same number of iterations as in the previous experiment to allow for a fair comparison.

4.3 Results and Discussion

Evaluation data Small Phoneme Set Earlier Nets	DP only	Highlevel Net (1 frame input)	TDNN highlevel (5 frame input)	TDNN highlevel (5 frame input & dynamic traing)
testing data (129)	81.4%	87.5%	89.1%	91.9%
training data (96)	74.0%	80.2%	89.6%	92.7%
testing+training (225)	78.2%	84.4%	89.3%	92.9%

Table 2: DP-TDNN: static vs. dynamic training

Table 2 shows the results for the four conditions outlined in the previous section. As can be seen the added higher level network improves results, as it smoothes and biases output firings appropriately. Higher level units overlooking a 50 msec (5 frames) window of phoneme spotting activations, increase this networks ability to appropriately smooth the output and eliminate spurious firings or drop-outs. Dynamic alignment during the training process does indeed increase performance on both training as well as testing data. This latter improvement is evidence, that integral optimization of sequential constraints and local, phonemic pattern recognition is a successful strategy towards improved and robust isolated

[4]Note again, that for computational efficiency, weights on only the higher level nets were actually changed here. Lower layer connections in the TDNNs were kept fixed here.

(and continuous) large vocabulary word recognition. Ultimately, recognition should not be limited by our insistence on preconceived notions and definitions of a phoneme. Rather phonemes should *emerge* in constraint satisfaction networks while they are learning to recognize words represented by phoneme symbol sequences.

5 Future Directions

In this section we have seen that Time-Delay Neural Networks can successfully be combined into vocabulary independent large vocabulary speech recognition systems. A number of open questions, however, remain to be addressed.

- We have seen that integral (dynamic) training leads to improved results over statically trained phoneme spotting nets. The experiments illustrating this improvement were, however, performed using only preliminary nets aimed at a subset of the phonemes. Integral training should be performed using the full set of improved phoneme spotting nets described in section 3.

- Errors should be propagated further down into the underlying TDNNs. So far (for computational efficiency) error backpropagation was performed only one layer deep during integral training.

- As error signal in our experiments we have so far only used the cumulative distortion over the course of a word, where distortion is measured by the mean square error between the target mask and the output activations of underlying phoneme spotting TDNNs. Alternative error measures should be explored. Some important observations should be made in this regard: Word level error should not (as it presently does) give equal weight to each time frame during the word. This leads to position dependence and penalizes networks for not producing phoneme target outputs even in regions that are not informative with respect to phoneme identity. To circumvent this problem, a basic TDNN [Waibel 87] integrates phoneme activations over time (in the second hidden layer), before passing the accumulated activation through a nonlinear decision function (a sigmoid function). In doing this, unimportant regions are deemphasized, while the network can more heavily rely on informative regions to achieve output criterion. When activations are integrated over an entire word, i.e., a sequence of regions with different targets, such a non-linear decision function should be introduced as well. Another important criterion for training should be word level optimization: Rather than trying to produce perfect phoneme spotting tracks, TDNNs should receive their error signal based on their ability to properly identify the target word. This could be done by making the

word level error a function of word identification performance, or discrimination from other words.

- Better models for duration and accent control should be applied. An initial simple model was introduced here, by way of a higher level net that utilizes a larger time-delay window to capture some duration dependent properties. Additional duration constraints could be introduced in the sequence manager (here the DP-matcher). Further, units could be introduced that use long time-delay windows that model long term trends in the energy or pitch contour to introduce additional prosodic constraints for word recognition.

- Phonemic transcriptions in the lexicon and access to them have to be improve or corrected.

- Alternate models for sequential management should be explored, in place of DP-matching. The assumption that adjacent states in a word are independent from each other is unsatisfactory for real speech. Similarly, a priori assumptions about the amount of warping allowable during DP-matching are unnatural and may impose artificial, rigid constraints.

- Extensions to continuous speech and speaker-independence should be explored. In view of speaker-independence, good initial results have recently been obtained using the CFM objective function [Hampshire 89a] and the Meta-Pie network [Hampshire 89b]. In continuous speech, a variety of different acoustic realizations are found for the same phoneme symbol. Networks have to be altered to embrace these acoustic variations. Most straight forward is to add tokens extracted from continuous speech to achieve proper generalization. Alternatively, it might be possible to partially retrain an existing net, and/or apply connectionist glue, to arrive at similar results with only limited additional training data and effort. Lastly, more probabilistic output measures (rather than the binary outputs currently in use) can be derived from existing nets that not only report the classification decision, but also the distances or probabilities from all other classification regions. While performance will always degrade if a system trained on one task is applied to another (such as going from one speaker, microphone to another or from isolated speech to continuous, etc.), results might degrade gracefully in this case, rather than abruptly.

References

[Bottou 88] Bottou, L-Y.
Reconnaissance de la Parole par Reseaux multi-couches.
In *Proceedings of Neuro-Nimes 88*. November, 1988.

[Bottou 89] Bottou, L., Fogelman-Soulie, F., Blanchet, P., Lienard, J.S.
Experiments with Time-Delay Networks and Dynamic Time Warping for Speaker Independent Isolated Digits Recognition.
In *Proceedings of the Eurospeech*. September, 1989.

[Elman 87] Elman, J.L. and Zipser, D.
Learning the Hidden Structure of Speech.
Technical Report, University of California, San Diego, February, 1987.

[Elman 88] J. L. Elman.
Finding Structure in Time.
Technical Report CRL Technical Report 8801, University of California, San Diego, 1988.

[Hampshire 89a] Hampshire, J. and Waibel, A.
A Novel Objective Function for Improved Phoneme Recognition Using Time-Delay Neural Networks.
In *Proceedings of the 1989 International Joint Conference on Neural Networks*. June, 1989.
(in review).

[Hampshire 89b] Hampshire, J.B. and Waibel, A.
The Meta-Pie Network: Building Distributed Knowledge Representations for Robust Pattern Recognition.
Technical Report CMU-CS-89-166, Carnegie-Mellon University, August, 1989.

[Hirai 89] Hirai, A. and Waibel, A.
Phoneme-Based Word Recognition by Neural Network -A Step Toward Large Vocabulary Recognition.
Technical Report, Carnegie Mellon University, August, 1989.

[Itakura 75] F.Itakura.
Minimum Prediction Residual Principle Applied to Speech Recognition.
IEEE Transactions on Acoustics, Speech, Signal Processing ASSP-23(1):67-72, February, 1975.

[McDermott 89] McDermott E., and Katagiri, S.
Phoneme Recognition Using Kohonen's Learning Vector Quantization.
Technical Report TR-I-00??, ATR Interpreting Telephony Research Laboratories, January, 1989.

[Sagisaka 87] Sagisaka, Y., Takeda, K., Katagiri, S. and Kuwabara, H.
Japanese Speech Database with Fine Acoustic-Phonetic Transcriptions.
Technical Report, ATR Interpreting Telephony Research Laboratories, May, 1987.

[Sakoe 78] H.Sakoe, S.Chiba.
Dynamic Programming Optimization for Spoken Word Recognition.
IEEE Transactions on Acoustics, Speech, Signal Processing ASSP-26(1):43-49, February, 1978.

[Watrous 87] Watrous, R.L., Shastri, L. and Waibel, A.H.
Learned Phonetic Discrimination Using Connectionist Networks.
In *European Conference on Speech Technology*, pages 377-380. Edinburgh, September, 1987.

[Wong 86] Wong, M.K. and Chun, H.W.
Towards a Massively Parallel System for Word Recognition.
In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 37.4.1-37.4.4. April, 1986.