Connectionist Large Vocabulary Speech Recognition

Alex Waibel[1]

School of Computer Science, Carnegie-Mellon University, Pittsburgh,PA 15213

Abstract: In this paper, the problem of large vocabulary word recognition is addressed from a connectionist perspective. The problem is not only of practical interest but also of scientific importance, since a workable solution must integrate pattern recognition under consideration of sequential, symbolic constraints. We have developed two large vocabulary word recognition systems based on different speech recognition philosophies. One of the systems exploits the power of neural networks in performing accurate classification, the other the power of producing good non-linear function approximation and signal prediction. We present each system's operation and evaluate its performance. Both achieved respectable recognition scores in excess of 90% correct for vocabularies of up to 5000 words. We suggest further avenues towards improvement of either system and in the process discuss the relative strengths of either approach.

Keywords: neural networkds, connectionism, time-delay neural networks, predictive neural networks

1 Introduction

Recognition of speech by machine has been a fascinating topic of research that has for many years given rise to some of the most innovative and exciting models. It has always been driven by a mix of intuitions relating to system design and engineering on one side and human cognitive modeling on the other. It has always drawn a great deal of ideas, motivation and inspiration from a desire to understand human communication, while imposing the realism of practical engineering constraints and comparative performance measures. Connectionist models or "neural networks" have recently attracted considerable (and renewed) attention in speech recognition as they provide speech scientists with a cognitively plausible model of speech processing while at the same time introducing a novel, yet realistic engineering solution to the problem. A number of initial designs have produced in a short time performance results that compared favorably or exceeded those obtained by traditional speech processing techniques [1, 2, 3, 4]. On the other hand, most of these experiments were limited to small tasks or subproblems of the speech recognition problem such as phoneme classification [1, 2, 5] or small vocabulary word recognition [6, 7, 8].

While these results are encouraging given those limited domains, the question remains to be answered if and how this technology may be used effectively for the design of whole speech understanding

[1] We gratefully acknowledge IEICE for permission to reprint this paper. It has originally appeared in the journal of the IEICE, Vol.J73-D-II, No.8 pp. 1122-1131, (Aug. 1990) by the same author.

NATO ASI Series, Vol. F 75
Speech Recognition and Understanding. Recent Advances
Edited by P. Laface and R. De Mori
© Springer-Verlag Berlin Heidelberg 1992

systems. Indeed, a common criticism argues that connectionist models are but good classifiers but cannot handle the temporal, sequential nature of speech. As such, connectionist models may be attractive only in limited domains or toy problems, but would scale poorly to large vocabulary speech understanding systems. Although this criticism has been valid for a number of initial simple networks, extensions that overcome these limitations have been proposed and are beginning to produce respectable results on larger problems as well.

In this paper we will describe current research activity that addresses the large vocabulary recognition problem. We present two large vocabulary word recognition systems that illustrate that neural networks 1.) can be used productively for large vocabulary speech recognition by way of classification but also by way of non-linear mapping and system identification 2.) neural networks can be integrated with connectionist as well as non-connectionist strategies to handle temporal, sequential processing to form chains of subword units, words and sentences.

2 The Large Vocabulary Word Recognition Problem

Early on connectionist word recognition experiments were carried out that have exploited the classification capabilities of neural nets by applying an entire word's coefficient matrix to the inputs of static full word networks with output units for each word to be classified. Good results were achieved, but the resulting systems required precise time alignment and a preprocessing stage that determines the endpoints of an input word, both unacceptable requirements in practice in the light of continuous speech, noise and varying speaking rates. Similarly limiting is the fact that only small vocabularies can be handled in this fashion, because network size and training time become prohibitively large and enrollment impractical with increasing vocabulary size.

To overcome the former first set of limitations, networks that model time, temporal distortion (warping) and/or shift-invariance internally have been proposed for small vocabulary recognition. Among them are techniques that integrate neural network based classification with traditional schemes for time alignment and sequence management, such as the Dynamic Neural Net (DNN) [8, 9], word level Time-Delay Neural Networks (TDNNs) [10, 11], hybrid neural net classifiers and Hidden Markov Models [12] and Neural Prediction Models [13]. Most of these models have been tested on small vocabularies (Japanese, French and English digits) and have achieved excellent performance results, but all used dedicated models for each vocabulary word and are in their basic forms not appropriate for large vocabulary recognition.

To extend these models to large vocabulary recognition subword units such as phonemes or syllables must be employed. Since such subword units are limited in number large vocabularies can be constructed as different sequences of these atomic subunits. In large vocabulary word recognition then the task is to identify the most likely sequence of phonetic units that make up a legal word (preferably without requiring segmentation in the process). Several models have been proposed that express sequential constraints in a connectionist framework alone [14, 15, 16, 17, 18]. Alternatively, combinations between the perceived strengths of neural networks at the pattern recognition level with the strengths of traditional methods at modeling sequences such as Hidden Markov Models, Viterbi Decoding, or Dynamic Programming have also been proposed. Such "hybrid approaches" have recently gained in popularity as they appear to offer immediate access to the best of both worlds.

In the following we d
examples of two different
and "prediction based mo

3 Classification Bas

Neural networks have be
as well as at the word lev
that recognize phoneme
phonemic output hypothe

3.1 Time-Delay Neura

One of our attempts in dc
been shown to produce e
to provide a non-linear nc
of precise temporal align
incorporate current activ
input. Fig.1 illustrates
consonants /b, d, g/ (see

Initial experimentatio
phoneme sets only (/
recognition [20] and rec
benefitted from modular
(Meta-Pi network [20]) to
focus of attention or rapi
classification experimer
improvement over result
overcome problems rel
abstractions in the hidd
efficiency and flexibility
networks [2].

3.2 Large Vocabulary

Based on a Japanese
speaker-dependent exp
of large vocabulary reco
output categories over n

[2]no assumptions as to the u

e but good classifiers but
ıectionist models may be
› large vocabulary speech
of initial simple networks,
re beginning to produce

rge vocabulary recognition
strate that neural networks
›f classification but also by
; can be integrated with
uential processing to form

that have exploited the
ınt matrix to the inputs of
results were achieved, but
stage that determines the
ight of continuous speech,
mall vocabularies can be
tively large and enrollment

nporal distortion (warping)
ıgnition. Among them are
hemes for time alignment
d level Time-Delay Neural
/ Models [12] and Neural
ularies (Japanese, French
sed dedicated models for
ıbulary recognition.

as phonemes or syllables
rge vocabularies can be
ary word recognition then
› a legal word (preferably
n proposed that express
ılternatively, combinations
level with the strengths of
ırbi Decoding, or Dynamic
tly gained in popularity as

In the following we describe two connectionist large vocabulary recognition systems. They are examples of two different recognition philosophies. We will refer to them as "classification based models" and "prediction based models".

3 Classification Based Models

Neural networks have been shown to implement excellent non-linear classifiers both at the phonetic level as well as at the word level. Large vocabulary systems can therefore be implemented by neural networks that recognize phonemes or parts of phonemes (states) and evaluate how well a sequence of their phonemic output hypotheses match the legal sequence of a word.

3.1 Time-Delay Neural Networks

One of our attempts in doing this is based on the Time-Delay Neural Network (TDNN). This network has been shown to produce excellent phoneme discrimination performance [1]. This network was developed to provide a non-linear non-parametric[2] pattern classifier that can spot features or phonemes independent of precise temporal alignment (shift-invariance property). The network is a multilayer network of units that incorporate current activations from lower layers as well as time-delayed versions of them (context) as input. Fig.1 illustrates a TDNN trained to perform the discrimination task between the voiced stop consonants /b, d, g/ (see [19] for a more detailed description of its operation).

Initial experimentation with this class of networks was performed speaker-dependently on small phoneme sets only (/b,d,g/ discrimination), but extensions to high performance multi-speaker recognition [20] and recognition of all phonemes were soon achieved. Both problems significantly benefitted from modular and incremental learning [20, 21, 2]. By using an integrating supernetwork (Meta-Pi network [20]) to decide on how to gate an appropriate mix of speaker specific network decisions, focus of attention or rapid adaptation to speaker specific classification can be achieved. In *multi*-speaker classification experiments this resulted in speaker-*de*pendent recognition rates - a significant improvement over results from speaker-independent training. Modularity could also be used effectively to overcome problems related to scaling, training time and generalization. By exploiting the featural abstractions in the hidden units of previously trained networks modular training allowed for greater efficiency and flexibility of design while achieving performance greater than or equal to non-modular networks [2].

3.2 Large Vocabulary Recognition by TDNN

Based on a Japanese large vocabulary isolated word database (5240 words) [22, 19, 1] a number of speaker-dependent experiments were carried out to improve the TDNN's performance, particularly in view of large vocabulary recognition [23]. For use in word recognition, speech is to be classified into phoneme output categories over running speech (in this case over entire words spoken in isolation). As the original

[2] no assumptions as to the underlying probability distributions need to be made

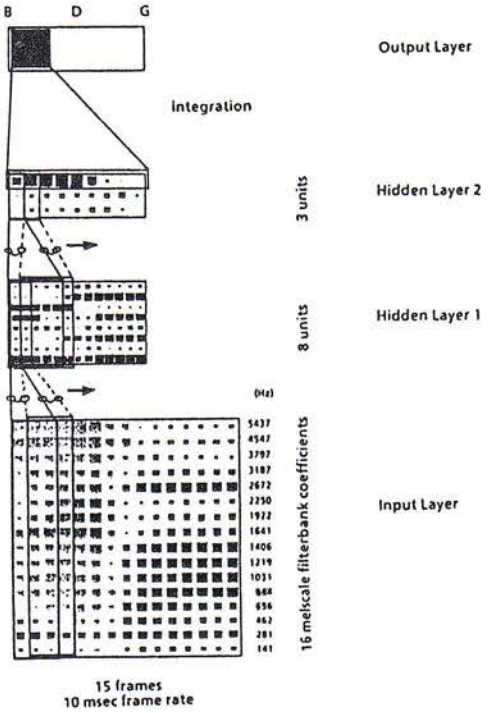


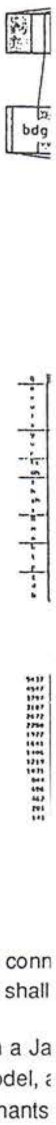
Figure 1. The TDNN architecture (/b,d,g/-task)

TDNNs were trained on excised phoneme tokens only, several modifications were desirable. First, the original excised phoneme training patterns were now artificially misaligned in time by various offsets. It more realistically simulates the absence of precise phoneme labels and segmentation. The resulting introduction of time alignment "noise" turned out not to decrease performance, but lead to noticable improvements instead, particularly for phoneme spotting. Training in this fashion improved generalization and enforced shift-invariant phoneme classification even in transitory regions between phonemes. The resulting phoneme spotting rates of the large scale TDNN's improved from 95.8% to 98.0% and more importantly, the false alarm rates[3] decreased from 62.2% to 23.2%[4]. The performance results of our earlier models and this improved model compared favorably with various other recognition strategies over the same data. For word recognition also a silence category was necessary which was added by modular design to the existing net [23]. Fig.2 shows the resulting large TDNN all-phoneme architecture. Fig.3 shows output activation patterns for the word "wata".

While good phoneme classification performance is indeed encouraging, this will have to be properly integrated and have to translate into good large vocabulary word recognition performance to advance the field. Mature speech recognition technology has already at its disposal a number of elegant techniques for this and similar word-level integration needs to be accomplished in a connectionist frame-work or in

[3] Presumably due to previously undefined transitory regions.

[4] All recognition tests were run on independent test data from the same speaker.



the form of hybrid conn
undertake and we shall

Using data from a Ja
phoneme level model, a
vowels, 18 consonants
for phoneme based w
sequence of phoneme
activations of a TDNN

'er

er 2

er 1

were desirable. First, the
time by various offsets. It
gmentation. The resulting
nce, but lead to noticable
on improved generalization
between phonemes. The
95.8% to 98.0% and more
performance results of our
recognition strategies over
hich was added by modular
oneme architecture. Fig.3

is will have to be properly
erformance to advance the
nber of elegant techniques
nectionist frame-work or in

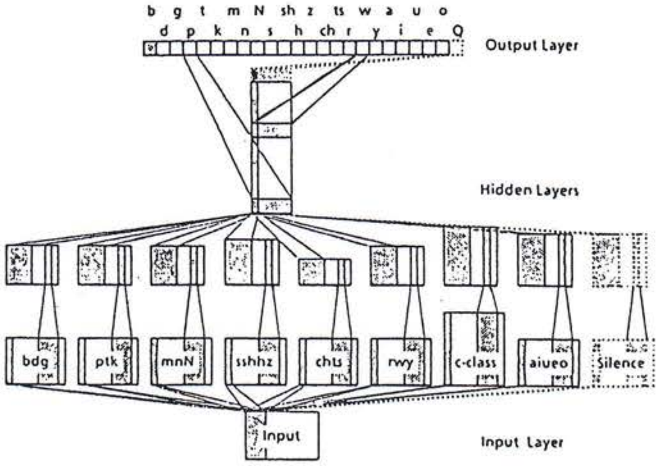


Figure 2. Modular TDNN used to spot all phonemes

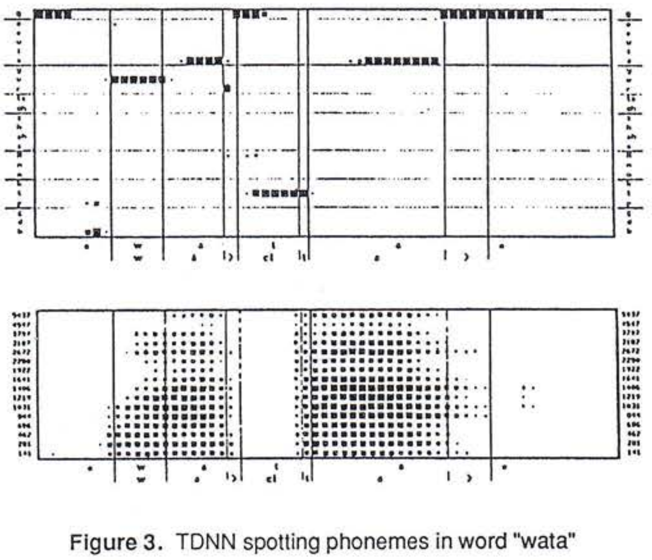


Figure 3. TDNN spotting phonemes in word "wata"

the form of hybrid connectionist/ non-connectionist system design. Neither is necessarily a trivial step to undertake and we shall desribe several successful initial attempts that have been proposed.

Using data from a Japanese isolated word database (as described above) and a TDNN as a front end phoneme level model, a hybrid large vocabulary recognition system was developed [23]. 24 phonemes (5 vowels, 18 consonants and silence) were spotted by shifting TDNNs across time providing the front end for phoneme based word recognition. To recognize a word, the overall likelihood of a word-specific *sequence* of phoneme activations needs to be estimated. To do so, we can approximate the output activations of a TDNN as representing the maximum a posteriori probabilities of a phoneme class given speech at a given time frame [24]. If each phoneme is viewed as a single state with an associated output probability, then a word likelihood can be calculated as the joint probability over all output probabilities over time. Assuming that all states are independent, a word likelihood would be given by the product of framewise outputs. A simple way of implementing this is to evaluate at each time frame the log activation of the output unit that corresponds to a legal phonemic state in the word and summing these log outputs over time. The correspondence between a given time frame and the current active phoneme node is performed by a Dynamic Time Warping (DTW) procedure.

An implementation of this procedure is described by Miyatake, Sawai, Minami and Shikano [23]. Here, a modular TDNN as described above was used, and only one state per phoneme was provided. An LR-parser provided top-down prediction of what set of phoneme transitions are legal to form legal words in the dictionary. For duration control each phoneme state was expanded to the average number of frames of that phoneme before DTW was carried out. Recognition experiments on various vocabulary sizes were undertaken with this system. All experiments were performed vocabulary independently[5] and on independent test data (phonemes not used for training). For a 500-word test vocabulary, first choice accuracy of 98% was achieved. For a large vocabulary of 5000 words, recognition rates as high as 92.6 were obtained. Second and fifth choice rates for the later vocabulary size were 97.6% and 99.1%, respectively, indicating that most confusions occurred among a small group of acoustically similar words (e.g., "itai" -> "ittai").

3.3 Extensions

The performance of the system described does indeed suggest that very high performance can be achieved, independent of training vocabulary and training context. Several problems, however, need to be overcome to further improve large vocabulary speech recognition systems.

Sequencing of Phoneme Internal Events: First, we have already noted that the TDNNs described above were all integrated as single phoneme states. While TDNNs can capture a variety of phoneme specific cues sequential ordering within a phoneme is only imposed within the reach of its fixed duration time-delays. Additional ordering between variably duration subphonemic states must be imposed in the context of word recognition. Variable or adaptive time-delays [25] could be used internally or a sequence of several states [12] per phoneme at its output. This should lead to better performance and duration modeling, particularly in continuously uttered poorly articulated speech.

Stochastic Modeling of Sequences: The most successful and popular approach to stochastic modeling of sequences is given by Hidden Markov Models (HMMs), where a phoneme is given by a stochastic sequence of states that can be linked together into words and from there on into sentences. At each of these levels (lexical, syntactic, etc.) constraints can be applied and probabilities estimated, and their joint probabilities (assuming they are independent) computed. A popular idea therefore is to use the strengths of neural networks at precise pattern classification in combination with the modeling of state sequences and time alignment found in HMMs.

Some of the earlier proposals at this were developed by Bourlard, Wellekens and Nelson [26, 24, 27]. In theoretical and experimental work they had shown that the outputs of a multilayer perceptron

[5] The phonemes used for training were extracted from words of a different vocabulary than the one used for testing.

(feedforward network) tra
estimates of the maximu
class. They have since
activations of a local mu
traditional HMM. Viterbi
states and to compute an

Several enhanceme
Bourlard [27] achieved si
outputs (the a posteriori
distributions in the trainin
usage of a cross-validati
given size, this can lead
on) new unseen data [2
yields a stopping criterio
third enhancement prop
CVT is akin to the segme
aims at integrated and se
(or state) segmentation a
best labeling of the inpu
outputs to correspond
performance, that are
sentence [27] and on cor

Research Directions
well for HMMs remain to
corrective training (at the
optimal HMM topology,
towards improved trainin
of local phonetic class
robustness.

An alternate exciting
constraints altogether i
models may relax some
potentially lead to further

4 Prediction Based

The connectionist mode
patterns or subpatterns.
approximates a bit pat
classifications, however,
Among them are non-

ith an associated output
r all output probabilities
given by the product of
frame the log activation
nming these log outputs
active phoneme node is

and Shikano [23]. Here,
eme was provided. An
egal to form legal words
the average number of
s on various vocabulary
ilary independently^5 and
vocabulary, first choice
on rates as high as 92.6
ere 97.6% and 99.1%,
coustically similar words

h performance can be
lems, however, need to

t the TDNNs described
e a variety of phoneme
ach of its fixed duration
must be imposed in the
internally or a sequence
rformance and duration

approach to stochastic
phoneme is given by a
e on into sentences. At
abilities estimated, and
a therefore is to use the
h the modeling of state

and Nelson [26, 24, 27].
a multilayer perceptron

sed for testing.

(feedforward network) trained by backpropagation from a mean square error may be considered to be estimates of the maximum a posteriori probabilities of a given input to belong to its corresponding output class. They have since built on this notion to construct Hidden Markov Model chains where the output activations of a local multilayer perceptrons (MLP) are used as output probabilities for the states in a traditional HMM. Viterbi aligment is performed to assign the framewise MLP firings to corresponding states and to compute an overall word output probability.

Several enhancements were subsequently proposed by several investigators. Morgan and Bourlard [27] achieved significant improvements in recognition performance, by normalizing their network outputs (the a posteriori probabilities) by their respective prior probabilities to eliminate a bias to uneven distributions in the training data. Another technique aimed at optimizing generalization performance is the usage of a cross-validation set. If only limited amounts of training data are available given a net of a given size, this can lead to overfitting to the training data and poor generalization to (poor performance on) new unseen data [27]. Use of an independent pseudo testing set (the cross-validation set) then yields a stopping criterion, that assures that a net is trained with optimal test-set performance in mind. A third enhancement proposed by several researchers is Connectionist Viterbi Training (CVT) [27, 12]. CVT is akin to the segmental k-means training procedure used for Hidden Markov model training [28] and aims at integrated and segmentation free word level training. The idea is to optimize a suitable phoneme (or state) segmentation *together* with the backpropagation network optimization. CVT iteratively finds the best labeling of the input (by way of Viterbi alignment), while the networks attempt to provide better outputs to correspond to these label. These techniques produced good word level recognition performance, that are beginning to compare favorably with other advanced HMMs on continuous sentence [27] and on connected digit [12] tasks.

Research Directions: A host of additional modifications and improvements that are known to work well for HMMs remain to be explored in the context of hybrid connectionist systems. Among them are corrective training (at the word level), choice of best input representation, transition probabilities, choice of optimal HMM topology, optimal neural network architecture, etc. Last not least, work is in progress towards improved training algorithms that generate more meaningful probability estimates at the outputs of local phonetic classification networks to improve *word level* discrimination and overall system robustness.

An alternate exciting research avenue is given by connectionist formalisms that represent sequential constraints altogether internally as connectionist modeling extensions [14, 15, 16, 17, 29, 18]. Such models may relax some of the limiting assumptions made by current recognition strategies and could potentially lead to further improvements in speech recognition system design.

4 Prediction Based Models

The connectionist models that we have discussed so far apply neural nets as classifiers of either word patterns or subpatterns. For classification, the input usually consists of a coefficient matrix and the output approximates a bit pattern representing the classification results. In addition to learning discrete classifications, however, neural networks can implement a variety of other constraint satisfaction tasks. Among them are non-linear function approximation, interpolation and prediction, which generate continuous real-valued output vectors. This can be exploited in speech for various signal mapping and coding applications, including noise suppression [30], speech code mapping [31] and non-linear signal prediction [32]. The use of neural networks as non-linear signal predictors in speech recognition has recently first been shown successfully in the "Neural Prediction Model" proposed by Iso and Watanabe [13] and the "Hidden Control Neural Network" proposed by Levin [33]. Both of these models have so far only been implemented for small vocabulary recognition tasks (i.e., digits), but have yielded high recognition performance speaker-independently. Extensions to large vocabulary recognition are also possible with this approach as we shall see in the following.

4.1 Recognition Using Small Vocabularies

The basic idea is illustrated in Fig.4. A two frame window of input coefficients is input into a multilayer feedforward net trained to produce at its output a frame of coefficients that is as close as possible to the next (future) speech frame. The distance between this predicted frame and the actual next speech frame can be measured as a prediction error or distortion and this distortion is used as error criterion for backpropagation training. Given a set of predictor networks one can imagine training each predictor for a separate region of an utterance. Each predictor net becomes specialized to best predict this portion of an utterance, such that the prediction error is likely to be lowest in these regions. A word is then represented by the *sequence* of predictor nets that best predicts the actual observed speech. Dynamic Programming is used as a mechanism to optimally apply each predictor sequentially over time to best approximate the actual signal. Fig.5 shows this alignment step based on the matrix of distances between actual speech frames and predicted frames. During training an alignment path is determined by Dynamic Programming. Each predictor is then trained to minimize the error between its output and the speech frames that it was assigned to predict according to the DP-alignment path. During recognition the word whose sequence of predictors minimizes the error between predicted frames and actual signal frames is chosen. Iso and Watanabe [13] used 10 mel scale cepstral coefficients and amplitude change as inputs to their networks. The number of predictors used depended on the utterance and ranged (for Japanese digits) between 9

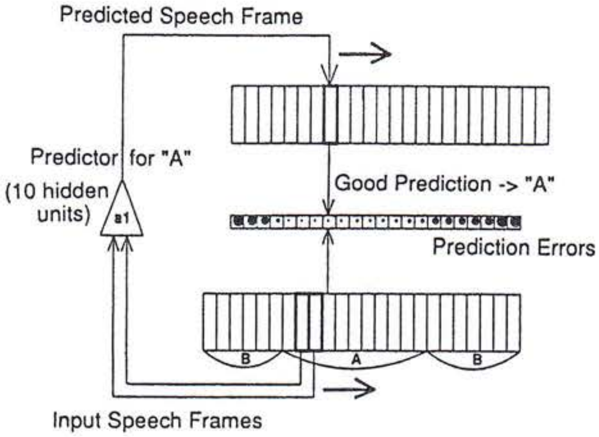


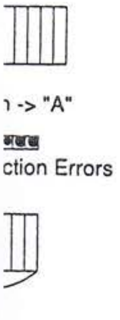
Figure 4. Modeling a phoneme by signal prediction

and 14. Each predictor
and 11 predicted output
speaker-independent iso
favorably with other tech
same data.

The model proposed t
it uses non-linear predi(
Prediction Model, howe
input flags or "control
progresses. Similar to
control the sequential s
12 deltacepstal parame
units turn on sequential
representation). Contr(

various signal mapping and
g [31] and non-linear signal
; in speech recognition has
el" proposed by Iso and
[33]. Both of these models
.e., digits), but have yielded
;abulary recognition are also

nts is input into a multilayer
; as close as possible to the
ıe actual next speech frame
used as error criterion for
training each predictor for a



ı

est predict this portion of an
A word is then represented
ːch. Dynamic Programming
me to best approximate the
ces between actual speech
l by Dynamic Programming.
e speech frames that it was
ıe word whose sequence of
frames is chosen. Iso and
as inputs to their networks.
Japanese digits) between 9

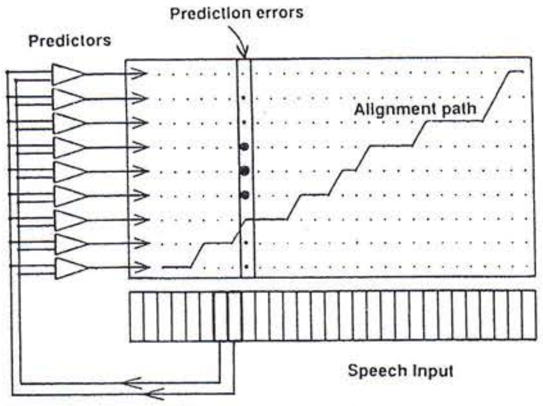


Figure 5. A Neural Prediction Model

and 14. Each predictor net has three layers, an input layer of two 11 coefficient frames, 9 hidden units and 11 predicted output coefficients. Excellent performance (0.2% error) was reported for a Japanese speaker-independent isolated digit recognition task uttered over telephone lines. This result comparėd favorably with other techniques (0.7% for the DNN [34, 8] and 1.1% for DP-matching [35]) tested on the same data.

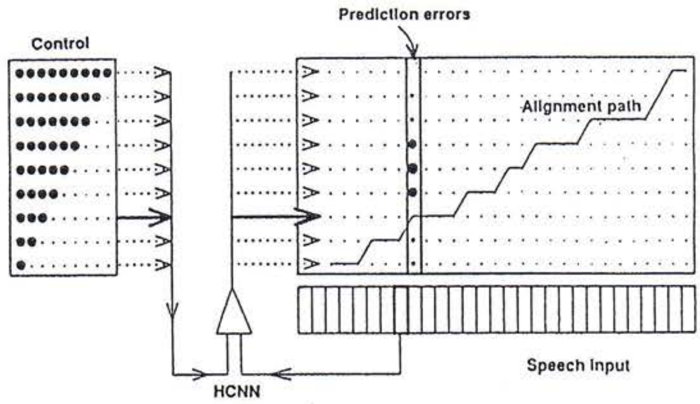


Figure 6. The Hidden Control Neural Network

The model proposed by Levin is similar to the one described above and is illustrated in Fig.6. As before it uses non-linear prediction by neural nets to measure a model's fit to the input data. Unlike the Neural Prediction Model, however, it uses only one single predictor for an entire word and a sequence of varying input flags or "control units" that switch the predictor into alternate modes of operation as time progresses. Similar to "counter nodes"(proposed for spelling correction [36]), these units are used to control the sequential state of the network. The predictor network used 24 speech inputs (12 cepstral and 12 deltacepstal parameters), 30 hidden units, 24 predicted outputs and 8 control input units. The control units turn on sequentially when appropriate and remain on as additional bits are activated ("thermometer" representation). Control transitions (the point at which a new bit is turned on) are determined by Viterbi

Alignment. During training, the Viterbi algorithm determines the state of control unit settings for each speech input frame and applies backpropagation learning to reduce prediction error according to this segmentation. The network was tested on connected digits from the TI-digit database (using male speakers only). Using independent test data but from the same speakers used in training, a word recognition rate of 99.3% was achieved.

4.2 Large Vocabulary Recognition

Large vocabulary word recognition using predictor networks is also possible. For use in large vocabulary recognition, words must here again be decomposed into subword units such as phones or syllables and an optimal model for these units must be trained. Recent work by Tebelskis and Waibel [37] has demonstrated that this can be done without the need for segmentation. In this work, time alignment and connection weights were optimised jointly and the weights of sets of network predictors corresponding to the same phoneme symbols were linked together (as in the TDNN). Experiments with the "Linked Predictive Neural Network" (LPNN) resulted in 94% recognition performance for speaker-dependent isolated word recognition over a database of 234 Japanese words and 90% over a 1000 word vocabulary. The data used in these experiments was given by a confusable subset of the data used for evaluation of the TDNN based system described in the previous section. Performance results on this particular subset were found to be comparable between the two systems.

The operation and training of the LPNN are shown in Fig.7. As before, a set of predictors is assigned to different portions of a word. Here these portions are defined to be phonemes and each occurrence of the same phoneme is modeled by the same set of three predictors. In Fig.7, for example, two words "BAB" and "ABA" may consist of the same phonemes in different order and position. Time alignment of the sequence of predictors is done as before, but all prediction errors assigned to the same phoneme (or portion thereof) train the same predictor net by way of a linkage pattern that defines the legal phoneme sequence of a word. A number of enhancements to this basic scheme have so far been found to be effective. A set of parallel predictors was added to each phoneme model to allow the LPNN to better represent alternate pronunciations and context dependencies. An assignment of each alternate was not predetermined, but the system selects the most suitable alternate based on the prediction errors produced by each alternate. During training the selected alternate is also reinforced by additional training while the others are not. In this fashion, the network automatically generates different models depending on context and pronunciation. A measurable performance improvement was obtained from this technique. Significant improvements were also obtained when phoneme pairs that are only distinguishable on the basis of duration (e.g., in Japanese: "k" vs. "kk") were represented by different sets of predictors. Fig.8 shows an example of processing in the LPNN for an input word "kashikoi". In the top panel, the original spectrogram is shown with 16 spectral coefficients per time frame and time moving from left to right. Underneath, the output predictions of the best predictors (as determined by DTW) at each time frame are displayed. The third panel shows ouput predictions for only one /i/-predictor(s). As can be seen prediction is best in the region corresponding to the final /i/, and degrades in other areas. The final display shows the distance marix obtained for each input frame and for each predictor linked into the word. Alignment is performed based on this matrix and the resulting labeling is shown at the input axis.

4.3 Extensions

To further enhance predi
addressed. The strength
for word level integration
suitably represent a word
with this approach is the
Fig.8 from the relative qu
word level integration, bu
also potentially more sen
alternate contexts or prc
Markov Models, such as
self-organizing principles

5 Conclusion

In this paper we have re
mere classification of sc
where constraints arisin
the additional constraint
limited this discussion to
included in complete lar

control unit settings for each
iction error according to this
l-digit database (using male
ers used in training, a word

For use in large vocabulary
ι as phones or syllables and
ɔelskis and Waibel [37] has
his work, time alignment and
predictors corresponding to
periments with the "Linked
ınce for speaker-dependent
ver a 1000 word vocabulary.
data used for evaluation of
ults on this particular subset

t of predictors is assigned to
and each occurrence of the
example, two words "BAB"
on. Time alignment of the
to the same phoneme (or
defines the legal phoneme
e so far been found to be
allow the LPNN to better
t of each alternate was not
on the prediction errors
orced by additional training
different models depending
was obtained from this
me pairs that are only
presented by different sets
word "kashikoi". In the top
ne frame and time moving
is determined by DTW) at
nly one /i/-predictor(s). As
degrades in other areas.
r each predictor linked into
ling is shown at the input

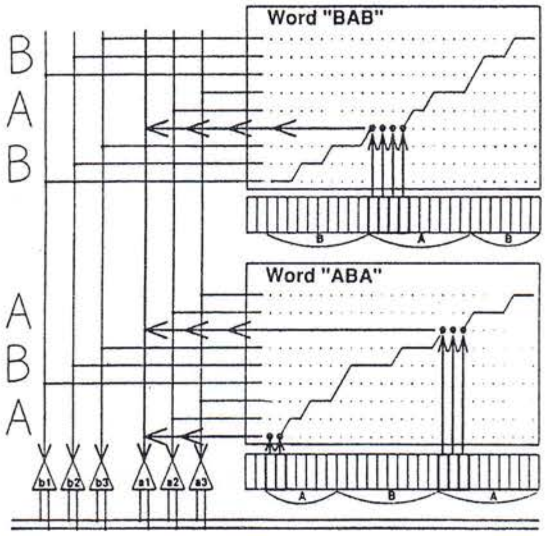


Figure 7. Training a Linked Predictive Neural Net

4.3 Extensions

To further enhance prediction based large vocabulary recognition, several current limitations have to be addressed. The strength of the model described here is that it inherently provides for simple mechanisms for word level integration and optimization. Optimization essentially proceeds top down, in an attempt to suitably represent a word's speech pattern given the phonetic sequence of the word. A possible problem with this approach is the apparent lack of discrimination at the speech pattern level as can be seen in Fig.8 from the relative quality of a single /i/-predictor applied to the entire utterance. This leads to good word level integration, but can result in poor acoustic-phonetic discriminability [38]. The representation is also potentially more sensitive to varying phonetic contexts [38], unless one provides alternate models for alternate contexts or pronunciations. This suggests enhancements similar to those applied to Hidden Markov Models, such as corrective training and context dependent phones. Alternatively, connectionist self-organizing principles could be attempted.

5 Conclusion

In this paper we have reviewed connectionist strategies applied to speech recognition. Reaching beyond mere classification of sound patterns, we have addressed the problem of large vocabulary recognition, where constraints arising from the classification of the underlying speech sounds must be interwoven with the additional constraints of sequentiality and lexical legality. We have on the other hand deliberately limited this discussion to the word level and not addressed sentence level issues that certainly have to be included in complete large vocabulary speech understanding systems (see [39, 40] for further discussion).

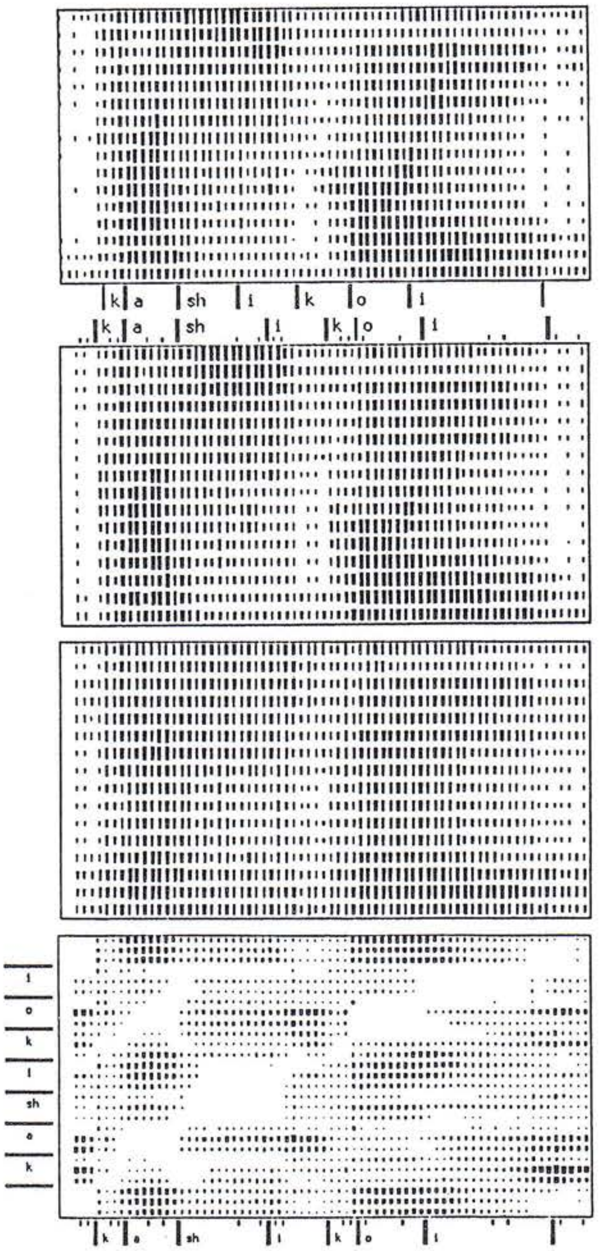


Figure 8. LPNN prediction for the word "kashikoi"
(See text for explanation)

We have developed two different connectionist large vocabulary systems, based on different underlying recognition philosophies. One is based on *classification*, the other on *prediction* of speech. Both

strategies achieved excelle
other. Interestingly, either
respective bottom-up or top
recognition philosophy are
humans' ability to use wha
phonetic distinctions. Our
machine will undoubtably c

1. Waibel, A., Hanaza
Time-Delay Neura
Processing, March

2. Waibel, A., Sawai,
Networks", *IEEE Tr*

3. Moore, R.K. and
Propagation Netwo

4. Robinson, A.J. and
Proceedings of nEu

5. McDermott, E., Iwa
HMM for Phoneme

6. Burr, D.J., "A Neur
and Cybernetics, O

7. Burr, D.J., "Expe
Transactions on Ac

8. Sakoe, H., Isotani
Recognition Using
Acoustics,Speech,

9. Isotani, R., Yoshid
Speech Recogniti
Technical Report,

10. Bottou, L-Y., "Rec
Nimes 88, Novemb

11. Bottou, L., Fogel
Networks and Dy
Proceedings of the

12. Franzini, M.A., Le
Continuous Spee
Signal Processing

13. Iso, K. and Wata
Model", *IEEE Inte*
1990.

ied on different underlying
ediction of speech. Both

strategies achieved excellent recognition performance and performed comparably with respect to each other. Interestingly, either approach displayed different areas of strength and weakness, related to their respective bottom-up or top-down recognition philosophies. While near-term enhancements using either recognition philosophy are being explored, one may wonder what kind of model may ultimately mimick humans' ability to use whatever constraints to recognize speech, be they high level pragmatic or fine-phonetic distinctions. Our search for an understanding of cognitive mechanisms and their realization by machine will undoubtably continue.

References

1. Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. and Lang K., "Phoneme Recognition Using Time-Delay Neural Networks", *IEEE, Transactions on Acoustics, Speech and Signal Processing*, March 1989.

2. Waibel, A., Sawai, H. and Shikano, K., "Modularity and Scaling in Large Phonemic Neural Networks", *IEEE Transactions on Acoustics, Speech, Signal Processing*, December 1989.

3. Moore, R.K. and Peeling, S.M., "Minimally Distinct Word-Pair Discrimination Using a Back-Propagation Network", *Computer, Speech and Language*, Vol. 3, No. 2, 1989, pp. 119-132.

4. Robinson, A.J. and Fallside, F., "A Dynamic Connectionist Model for Phoneme Recognition", *Proceedings of nEuro'88*, IEE, 1988.

5. McDermott, E., Iwamida, H., Katagiri, S. and Tohkura, Y., *Shift-Tolerant LVQ and Hybrid LVQ-HMM for Phoneme Recognition,* Morgan Kaufmann, 1990.

6. Burr, D.J., "A Neural Network Digit Recognizer", *IEEE International Conference on Systems, Man, and Cybernetics*, October 1986.

7. Burr, D.J., "Experiments on Neural Net Recognition of Spoken and Written Text", *IEEE Transactions on Acoustics, Speech; Signal Processing*, July 1988, pp. 1162-1168.

8. Sakoe, H., Isotani, R., Yoshida, K., Iso, K., and Watanabe, T., "Speaker-Independent Word Recognition Using Dynamic Programming Neural Networks", *IEEE International Conference on Acoustics,Speech, and Signal Processing*, May 1989, pp. 29-32.

9. Isotani, R., Yoshida, K., Iso, K., Watanabe, T. and Sakoe, K., "Dynamic Neural Network --- A New Speech Recognition Model Based on Dynamic Programming and Neural Network", *IEICE Technical Report*, September 1988.

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

11. 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", *Proceedings of the Eurospeech*, September 1989.

12. Franzini, M.A., Lee, K.F.,Waibel,A.H., "Connectionist Viterbi Training: A New Hybrid Method for Continuous Speech Recognition", *IEEE International Conference on Acoustics,Speech, and Signal Processing*, April 1990.

13. Iso, K. and Watanabe, T., "Speaker-Independent Word Recognition Using A Neural Prediction Model", *IEEE International Conference on Acoustics,Speech, and Signal Processing*, IEEE, April 1990.

14. Wong, M.K. and Chun, H.W., "Towards a Massively Parallel System for Word Recognition", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1986, pp. 37.4.1-37.4.4.

15. Lippmann R.P. and Gold, B., "Neural-Net Classifiers Useful for Speech Recognition", *IEEE International Conference on Neural Networks*, June 1987.

16. J. L. Elman, "Finding Structure in Time", Tech. report CRL Technical Report 8801, University of California, San Diego, 1988.

17. Bridle, J.S., "Alpha-Nets: A Recurrent Neural Network Architecture with a Hidden Markov Model Interpretation", *Speech Communication*, 1990, (to appear)

18. Young, S.J., "Competitive Training: A Connectionist Approach to the Discriminative Training of Hidden Markov Models", Tech. report CUED/F-INFENG/TR.41, Cambridge University, March 1990.

19. Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. and Lang K., "Phoneme Recognition Using Time-Delay Neural Networks", Tech. report TR-1-0006, ATR Interpreting Telephony Research Laboratories, October 1987.

20. Hampshire, J. and Waibel, A., "The Meta-Pi Network: Connectionist Rapid Adaptation for High-Performance Multi-Speaker Phoneme Recognition", *IEEE International Conference on Acoustics,Speech, and Signal Processing*, IEEE, April 1990.

21. Waibel, A., "Modular Construction of Time-Delay Neural Networks for Speech Recognition", *Neural Computation, MIT-Press*, March 1989.

22. Sagisaka, Y., Takeda, K., Katagiri, S. and Kuwabara, H., "Japanese Speech Database with Fine Acoustic-Phonetic Transcriptions", Tech. report, ATR Interpreting Telephony Research Laboratories, May 1987.

23. Miyatake, M., Sawai, H., Shikano, K., "Integrated Training for Spotting Japanese Phonemes Using Large Phonemic Time-Delay Neural Networks", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 1990.

24. Bourlard, H. and Wellekens, C.J., "Speech Pattern Discrimination and Multilayer Perceptrons", *Computer, Speech and Language*, Vol. 3, 1989, pp. 1-19.

25. U. Bodenhausen, "The Tempo Algorithm: Learning in a Neural Network with Adaptive Time-Delays", *Proceedings of the IJCNN*, IJCNN, January 1990, pp. 597-600.

26. Bourlard, H. and Wellekens, C.J., "Links between Markov Models and Multilayer Perceptrons", *Advances in Neural Network Information Processing Systems*, Morgan Kaufmann, 1988.

27. N. Morgan and H. Bourlard, "Continuous Speech Recognition Using Multilayer Perceptrons with Hidden Markov Models", *IEEE International Conference on Acoustics,Speech, and Signal Processing*, IEEE, April 1990, pp. 26.S8.1.

28. Rabiner, L.R, Wilpon, J.G. and Juang, B.H., "A Segmental K-Means Training Procedure for Connected Word Recognition", *AT&T Technical Journal*, May 1986.

29. Niles, L.T. and Silverman, H.F., "Combining Hidden Markov Model and Neural Network Classifiers", *IEEE International Conference on Acoustics,Speech, and Signal Processing*, IEEE, April 1990, pp. 417-420.

30. Tamura, S. and Waibel A., "Noise Reduction Using Connectionist Models", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1988, pp. S12.7.

31. Atal, B., "Non-Linear Mapping between Speech Codes", Personal Communication

32. Lapedes A. and Far
System Modeling",

33. Levin, E., "Speech
of the International

34. Sakoe, H., "Dynami
Programming and N

35. H.Sakoe, S.Chiba,
Transactions on Ac
43-49.

36. Kukich, K., "Back
Conference on Neu

37. Tebelskis, J. and
Networks", *IEEE Ir*
April 1990.

38. Endo, T., Tamura,
Models", Tech. re
1989.

39. Waibel, A. and Lee
Mateo, CA, 1990.

40. Furui, S. and Sondl
New York, NY, 199

for Word Recognition", *IEEE*
rocessing, April 1986, pp.

Speech Recognition", *IEEE*

ıl Report 8801, University of

vith a Hidden Markov Model

e Discriminative Training of
ımbridge University, March

honeme Recognition Using
reting Telephony Research

Rapid Adaptation for High-
rnational Conference on

for Speech Recognition",

Speech Database with Fine
ng Telephony Research

Japanese Phonemes Using
Conference on Acoustics,

ıd Multilayer Perceptrons",

work with Adaptive Time-
ı.

ıd Multilayer Perceptrons",
Kaufmann, 1988.

Multilayer Perceptrons with
stics,Speech, and Signal

ns Training Procedure for

del and Neural Network
Signal Processing, IEEE,

odels", *IEEE International*
p. S12.7.

nunication

32. Lapedes A. and Farber R., "Nonlinear Signal Processing Using Neural Networks; Prediction and System Modeling", Tech. report LA-UR-87-2662, Los Alamos National Laboratory, 1987.

33. Levin, E., "Speech Recognition Using Hidden Control Neural Network Architecture", *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, IEEE, April 1990.

34. Sakoe, H., "Dynamic Neural Network --- A New Speech Recognition Model Based on Dynamic Programming and Neural Network", *IEICE Technical Report*, December 1987.

35. H.Sakoe, S.Chiba, "Dynamic Programming Optimization for Spoken Word Recognition", *IEEE Transactions on Acoustics, Speech, Signal Processing*, Vol. ASSP-26, No. 1, February 1978, pp. 43-49.

36. Kukich, K., "Back-Propagation Topologies for Sequence Generation", *IEEE International Conference on Neural Networks*, 1988, pp. 301-308.

37. Tebelskis, J. and Waibel, A., "Large Vocabulary Recognition Using Linked Predictive Neural Networks", *IEEE International Conference on Acoustics,Speech, and Signal Processing*, IEEE, April 1990.

38. Endo, T., Tamura, S. and Nakamura, M., "Phoneme Recognition Using Neural Prediction Models", Tech. report TR-I-0107, ATR Interpreting Telephony Research Laboratories, August 1989.

39. Waibel, A. and Lee, K.F., *Readings in Speech Recognition,* Morgan Kaufmann Publishers, San Mateo, CA, 1990.

40. Furui, S. and Sondhi, M.M., *Advances in Acoustics and Speech Processing,* Marcel Dekker, Inc., New York, NY, 1990.