Speech Recognition and Understanding
Recent Advances, Trends and Applications

Edited by
Pietro Laface  Renato De Mori

NATO ASI Series

Series F: Computer and Systems Sciences, Vol. 75
Speech Recognition and Understanding
Recent Advances, Trends and Applications

Edited by
Pietro Lafface
Dipartimento di Automatica e Informatica
Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Torino, Italy

Renato De Mori
School of Computer Science
3480 University St., Montreal, Quebec H3A 2A7, Canada

Springer-Verlag
Berlin Heidelberg New York London Paris Tol
Hong Kong Barcelona Budapest
Published in cooperation with NATO Scientific Affairs Divis
Connectionist Large Vocabulary Speech Recognition

Alex Waibel

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 networks, 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
tems. Indeed, a common criticism argues that connectionist models are but good classifiers but not handle the temporal, sequential nature of speech. As such, connectionist models may be active 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, ansions that overcome these limitations have been proposed and are beginning to produce pectable 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 can be used productively for large vocabulary speech recognition by way of classification but also by 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 tions of subword units, words and sentences.

The Large Vocabulary Word Recognition Problem

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

To overcome the former first set of limitations, networks that model time, temporal distortion (warping) d/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 d 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 edition 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 ust be employed. Since such subword units are limited in number large vocabularies can be nstructed as different sequences of these atomic subunits. In large vocabulary word recognition there task is to identify the most likely sequence of phonetic units that make up a legal word (preferably ithout 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 of the perceived strengths of neural networks at the pattern recognition level with the strengths of additional 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 ey appear to offer immediate access to the best of both worlds.
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-dependent 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 noticeable 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 or this and similar word-level integration needs to be accomplished in a connectionist framework 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 connectionist/non-connectionist system design. Neither is necessarily a trivial step to undertake and we shall describe 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
ach at a given time frame [24]. If each phoneme is viewed as a single state with an associated output
ability, then a word likelihood can be calculated as the joint probability over all output probabilities
\time. Assuming that all states are independent, a word likelihood would be given by the product of
wise outputs. A simple way of implementing this is to evaluate at each time frame the log activation
output unit that corresponds to a legal phonemic state in the word and summing these log outputs
\time. The correspondence between a given time frame and the current active phoneme node is
formed by a Dynamic Time Warping (DTW) procedure.

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

Extensions

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

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

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

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

5The phonemes used for training were extracted from words of a different vocabulary than the one used 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 alignment 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
nous real-valued output vectors. This can be exploited in speech for various signal mapping and 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 so far only been implemented for small vocabulary recognition tasks (i.e., digits), but have yielded recognition performance speaker-independently. Extensions to large vocabulary recognition are also possible with this approach as we shall see in the following.

**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 (future) speech frame. The distance between this predicted frame and the actual next speech frame is measured as a prediction error or distortion and this distortion is used as error criterion for propagation 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 was 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 signed 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 Atanabe [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.
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 compared favorably with other techniques (0.7% for the DNN [34, 8] and 1.1% for DP-matching [35]) tested on the same data.

Figure 5. A Neural Prediction Model

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 delta cepstral 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
ignment. During training, the Viterbi algorithm determines the state of control unit settings for each 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 cognition rate of 99.3% was achieved.

2 Large Vocabulary Recognition

Large vocabulary word recognition using predictor networks is also possible. For use in large vocabulary cognition, 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 optimized 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 confusible 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 reduced 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 om 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 output 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 matrix 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 xis.
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 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 mimic 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 undoubtedly continue.

References


