The Development of a Speech Recognition Software Application Using HMM, ANN and LPC Models to Minimize Error

Gokhan Ozkartal

Girne American University, Cyprus

E-mail: drnexusgokhan@gmail.com

Received: April 13, 2013 Accepted: April 25, 2013 Published: July 1, 2013

doi:10.5296/jmr.v5i3.3521 URL: http://dx.doi.org/10.5296/jmr.v5i3.3521

Abstract

The purpose of this application is to develop speech recognition software applications and test it. The developed speech recognition system for human language solves for processing the voice input and answering the queries of the user. This system begins with an introduction to Speech Recognition Technology then it explains how it works, and the level of accuracy that can be expected.

The system is considered for the possible applications of speech technology, including every single task related with the action of the human voice. In this sense, the application fields can vary from speech production, storage, transmission and recognition processes.

This research pertains to discuss the theories and models on speech communication and more importantly the use of technology for speech recognition. Several authors' views are discussed but a great emphasis is given to the *Hidden Markov Models (HMM)*, *Artificial Neural Network (ANN)* and *Liner Predictive Coding (LPC).* The aim of the research is to test these models and outline voice recognition algorithms which are important in improving the voice recognition performance.

The application is tested with speakers used to represent different tones, emotions and lengths in metres as a sample of relationships methodology, the reports the system permits are used in the results analysis.

Keywords: Applications, Human language, System

Introduction

Speech has been a natural mode of communication since mankind; all the relevant skills are learnt during early childhood, without instruction, and we continue to rely on speech communication throughout our lives. However, to date technology is playing an important role in developing communication further to make life more efficient in our everyday lives, although speech appears to come so naturally to mankind it is actually a complex phenomenon, evident from this research.

What is a common debate are the speech acoustic and temporal types of variability. The research takes an incremental approach to this problem. Of the two types of variability in speech acoustic and temporal the former is more naturally posed as a static pattern matching problem that is amenable to neural networks; therefore neural networks for acoustic modelling is utilised, while the research relies on conventional Hidden Markov Models for temporal modelling. The research thus represents an exploration of the space of *NN-HMM hybrids*. The author explores two different ways to use neural networks for acoustic modelling, namely *prediction* and *classification* of the speech patterns. An extensive series of experiments that were performed to optimize the networks for word recognition accuracy is presented to show that a properly optimized NN-HMM hybrid system based on classification networks can outperform other systems under similar conditions. Finally, the research argues that hybrid NN-HMM systems offer several advantages over pure HMM systems, including better acoustic modelling accuracy, better context sensitivity, more natural discrimination, and a more economical use of parameters.

Theoretical Background

Constraints in Speech Recognition

A common agreement between researchers (Doddington, 1989, Kimura, 1990, Miyatake, 1990, Hild & Waibel 1993) is that speech generation and therefore recognition is a very complex problem. The human vocal tract and articulators are biological organs with nonlinear a property, whose operation is not just under conscious control but also affected by factors ranging from gender to upbringing to emotional state. Therefore applying technology to an already complex area can be very difficult. For example, variability's such as vocalizations can vary widely in terms of the speaker's accent, pronunciation, articulation, roughness, nasality, pitch, volume, and speed; moreover, during transmission, or irregular speech patterns can be further distorted by background noise and echoes, as well as electrical characteristics such as using a telephone or other electronic equipment.

Rabiner and Juang (1993) in their research provide a theoretical and accurate description of the basic knowledge and ideas that constitute a modern system for speech recognition by a machine. They discuss characterization of the speech signal; signal processing and analysis methods for speech recognition, with particular notion to patterns and differences that exist in models such as hidden Markov models. However, the main argument appears to be that the current state of the art in speech recognition lies in the conditions under which it is evaluated: under sufficiently narrow conditions almost any system can attain human-like accuracy, but it's much harder to achieve good accuracy under general conditions such as background noise, speech style and vocabulary.

Neural networks can form the basis for a general purpose speech recognition system, and that neural networks offer some clear advantages over conventional techniques. Many biological details are ignored in simplified models. For example, biological neurons produce a sequence of pulses rather than a stable activation value; there exist several different types of biological neurons. Plutchik (1994) states that physical geometry can affect their computational behaviour; they operate asynchronously, and have different cycle times; and their behaviour is affected by hormones and other chemicals. Although it may be necessary for modelling the brain's behaviour, the simplified model has enough computational power to support interesting research. Connectionism, or the study of artificial neural networks, was initially inspired by neurobiology, but it has since become a very interdisciplinary field, spanning computer science, electrical engineering, mathematics, physics, psychology, and linguistics as well. Several properties are closely linked to the ideas such as **Trainability** for which can be taught to form associations between any input and output patterns, **Generalization** networks don't just memorize the training data; rather, they learn the underlying patterns, so they can generalize from the training data to new examples. **Nonlinearity** networks can compute nonlinear, nonparametric functions of their input, enabling them to perform arbitrarily complex transformations of data. This is useful since speech is a highly nonlinear process. **Robustness** networks are tolerant of both physical damage and noisy data; in fact noisy data can help the networks to form better generalizations. This is a valuable feature, because speech patterns are generally noisy. **Uniformity** networks offer a uniform computational paradigm which can easily integrate constraints from different types of inputs and **Parallelism** networks are highly parallel in nature, so they are well-suited to implementations on massively parallel computers. Neural networks are usually used to perform static pattern recognition, that is, to statically map complex inputs to simple outputs, such as an N-ary classification of the input patterns. Moreover, the most common way to train a neural network for this task is via a procedure called *back propagation* (Rumelhart *et al*, 1986), whereby the network's weights are modified in proportion to their contribution to the observed error in the output unit activations (relative to desired outputs). To date, there have been many applications of neural networks trained by back propagation.

Speech recognition is also known as *automatic speech* recognition or *computer speech recognition* converts spoken words to text (Kranzlmuller, Reitinger, Hackl, & Volkert, 2001). The term "voice recognition" is sometimes used to refer to recognition systems that must be trained to a particular speaker as is the case for most desktop recognition software. Recognizing the speaker can simplify the task of translating speech, as there could be a mismatch caused by biological details as mentioned earlier. Although signal processing technologies show promise in leading to robust systems, the fundamental method for improving robustness is to understand the reason of the mismatch between the training data used in system development and testing data gathered in the real environment. The main causes of acoustic variations resulting from the speech production processes undertaken from other researchers work (Kranzlmuller, Reitinger, Hackl & Volkert, 2001), suggest that microphone, noise, emotions (stress), speaker quality and distortion are the main reasons why often why errors occur. Therefore, very often words may not be recognized.

One of the oldest and most important algorithms in speech recognition is the *Dynamic Time Warping* algorithm (Vintsyuk 1971, Itakura 1975, Sakoe & Chiba 1978). The simplest way to recognize an *isolated* word sample is to compare it against a number of stored word templates and determine which the “best match” is. This goal is complicated by a number of factors. First, different samples of a given word will have somewhat different durations. This problem can be eliminated by simply normalizing the templates and the unknown speech so that they all have an equal duration. However, another problem is that the rate of speech may not be constant throughout the word; in other words, the optimal alignment between a template and the speech sample may be nonlinear. Dynamic Time Warping (DTW) is an efficient method for finding this optimal nonlinear alignment.

Models for minimising error

Neural networks are usually used to perform static pattern recognition, that is, to statically map complex inputs to simple outputs, such as an N-ary classification of the input patterns. Moreover, the most common way to train a neural network for this task is via a procedure called *back propagation* (Rumelhart *et al*, 1986), whereby the network’s weights are modified in proportion to their contribution to the observed error in the output unit activations (relative to desired outputs). To date, there have been many applications of neural networks trained by back propagation. Some of the models are *NET talk* (Sejnowski & Rosenberg, 1987) is a neural network that learns how to pronounce English text, *Neurogammon* (Tesauro 1989) is a neural network that learns a winning strategy for Backgammon, *ALVINN* (Pomerleau 1993) is a neural network that learns how to drive a car and *Handwriting recognition* (Le Cun et al, 1990) is based on neural networks and has been used to read ZIP codes on US mail envelopes. Because theses examples are made-up or artificaial we can also call these Artificial Neural Networks (ANN).

Speech recognition, of course, has been another proving ground for neural networks. Althugh researchers such as Watrous (1988), Franzini et al, 1989 and Waibel *et al,* (1989) have made much progress in basic tasks as voiced/unvoiced discrimination, speech recognition still remains to be seen whether neural networks could support a large vocabulary, speaker independent, continuous speech recognition system, and also biological details have not been taken in to consideration with those models mentioned above.

As mentioned above DTW is one method particularly for optimal nonlinear alignment. An instance of the general class of algorithms known as *dynamic programming*. Its time and space complexity is merely linear in the duration of the speech sample and the vocabulary size. The algorithm makes a single pass through a matrix of frame scores while computing locally optimized segments of the global alignment path. (See Figure 2.6.) If *D*(*x*,*y*) is the Euclidean distance between frame *x* of the speech sample and frame *y* of the reference template, and if *C*(*x*,*y*) is the cumulative score along an optimal alignment path that leads to (*x*,*y*), then *C x y*= MIN *C x* – 1,*y C x* – 1,*y* – 1 *C x y* – 1 + *D x y*: (Adapted from: Ackley, D., Hinton, G., & Sejnowski, T., 1985).

The use of Hidden Markov Models (HMMs) for speech recognition has become predominant in developing speech recognition models. Its popularity is highly related to the inherent statistical (mathematically precise) framework (Bridle, 1990). The model is recognized as being easy to utilize, its accurate training algorithms for estimating the parameters of the models from finite training sets of speech data; the flexibility of the resulting recognition system in which one can easily change the size, type, or architecture of the models to suit particular words, sounds, and so forth; and the ease of implementation of the overall recognition system. Therefore the models have been the most flexible and successful approach to speech recognition to date.

HMM is a powerful statistical tool for modelling generative sequences that can be characterised by an underlying process generating an observable sequence. HMMs have found application in many areas interested in signal processing, and in particular speech processing, but have also been applied with success to low level natural language processing (NLP) tasks such as part-of-speech tagging, phrase chunking, and extracting target information from documents.

According to Waibel and Lee (1990) when an HMM is applied to speech recognition, the states are interpreted as acoustic models, indicating what sounds are likely to be heard during their corresponding segments of speech; while the transitions provide temporal constraints, indicating how the states may follow each other in sequence. Therefore, this can be considered as a structure, because speech always goes forward in time, transitions in a speech application always go forward. Fig. 1 illustrates how states and transitions in an HMM can be structured, in order to represent phonemes, words, and sentences.

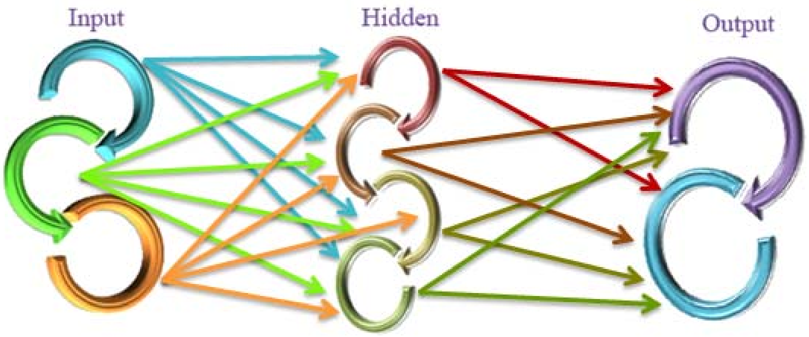

Figure 1. A hierarchically structured HMM

(Adapted from Waibel and Lee, 1990)

According to Tebelskis (1990) there are three basic algorithms associated with Hidden Markov Models these can be outlined as the *forward algorithm*, useful for isolated word recognition, the *Viterbi algorithm*, useful for continuous speech recognition; and the *forward-backward algorithm*, useful for training an HMM.

Formally, an HMM consists of the following elements:

{*s*} = A set of states.

{*aij*} = A set of transition probabilities, where *aij* is the probability of taking the transition from state *i* to state *j*.

{*bi(u)*} = A set of emission probabilities, where *bi* is the probability distribution over the acoustic space describing the likelihood of emitting1 each possible sound

u while in state *i*.

Since *a* and *b* are both probabilities, they must satisfy the following properties:

$$a_{ij} \geq 0, \quad b_i(u) \geq 0, \quad \forall i,j,u$$

$$\sum_j a_{ij} = 1, \quad \forall i$$

$$\sum_u b_i(u) = 1, \quad \forall i$$

Figure 2. (Cited from: Bridle, 1990 p.83-92)

Therefore, as outlined above (Fig. 2.) in which ***a*** and ***b*** depend only on the current state, independent of the previous history of the state sequence, limits the number of trainable parameters and makes the training and testing algorithms very efficient, proving that HMMs are useful for speech recognition.

The application and test

Speech recognition applications include voice user interfaces such as voice dialling "Call home", call routing "I would like to make a collect call", Home automation appliance control, search find a podcast where particular words were spoken, simple data entry entering a credit card number preparation of structured documents a radiology report, speech-to-text processing word processors or emails, and aircraft usually termed Direct Voice Input. However, as mentioned there are many difficulties associated to developing speech recognition programmes even with the more sophisticated models of Neural Systems and HMM. The proposed methodology is to develop a programme for speech recognition that takes in to consideration those deficiencies outlined and test them. Therefore the aim of this research was to minimise errors that was previously outlined.

The application training process permitted a command list to generate an active command list report that provides information about the number of speakers with an overview (overdue) fine balance, the total of overdue fine using software, the total number of words, and list of users which were selected as twenty times. The systems permit the user to remove words from the active command list, as systems that are checked out will automatically be checked in before being removed from the command list. It also permits the user to generate a circulation report that provides information about the number of active users, the volume of open words, and the most frequently close words. Hence, providing a report in conclusion compares results.

As mentioned above, voice recognition works are based on the promise that a person voice exhibits characteristics are unique to different speakers. The signal during training and testing session can be greatly different due to many factors such as people voice change with time, health condition (e.g. the speaker has a cold), speaking rate and also acoustical noise and variation recording environment via microphone. This training therefore provided opportunity for the analysis of sensitivity and specificity for the relationships between a software system and its users. Speakers used were selected to represent different tones, emotions and lengths in metres. Table 1 gives detailed information of recording and training sessions that took place with an indication of the overall methodology proposed.

Table 1. Training

Process	**Description**
1) Speaker	One Female One Male
2) Tools and Equipment	Mono microphone Gold Wave software Parallel Port Lamp Kit
3) Venue	Laboratory
4) Utterance	Twice each of the following words: On /Off Microphone Input Volume Increase Input Volume Decrease Channel Sterio
5) Sampling Frequency, fs	16000Khz
6) Feature computational	39 double delta mel-scale frequency cepstral coefficient.

As indicated in table 1, one male and one female was selected as the users of the application during the test. This was selected to include the variations in human sound and emotions as mentioned. The training also took place with distance and volume sampling as outlined in table 2. The test was run 20 times in order to include a measurable amount to compare. This sampling and methodology was considered feasible to make the research more repost.

Table 2. Distance and Volume Sampling

Distance	Volume
1-2 Meter	whisper conversation loud
1 Meter	whisper conversation loud
4 Meter	whisper conversation loud

Results of the test

To assess the specificity, the command “Open and Close” was used, because it is similar to “activate”. The Open and Close command was repeated three times and the number of false detections recorded. This test was performed for three different distances between the speaker and the microphone, and for three different levels of volume (whisper, conversation, and loud), whilst Figure 2 shows the flowchart for overall voice recognition process.

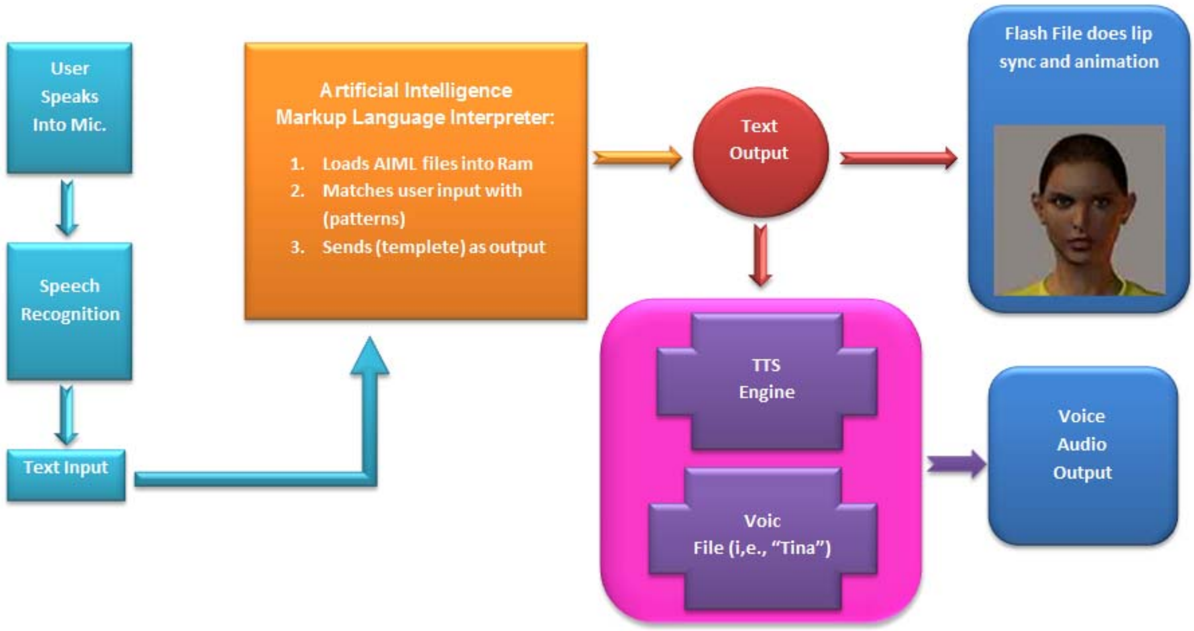
User Speaks Into Mic.
Speech Recognition
Text Input
Artificial Intelligence Markup Language Interpreter:
1. Loads AIML files into Ram
2. Matches user input with (patterns)
3. Sends (templete) as output
Text Output
Flash File does lip sync and animation
TTS Engine
Voic File (i,e., “Tina”)
Voice Audio Output

Figure 2. Flow chart of the system software

The speakers maintained relationships with this software. Users exist within the system and represent specific ways the system can be used. This software can participate in three types of relationships: tone, emotion and length. The *relationships* are the link between users. User often list of common characteristics. Suppose you identified a second usage: Check In word. This second usage commands certain features with the out of use and usage. Both using perform a transaction that affects the words inventory. A generalization allows you to represent this shared functionality in a third using (the Perform Transaction usage) and inherit its functionality in both the Check In words and the Check Out words usage. This relationship allowed one user to include the functionality of another. Before one word can be checked out, the system verified that the speaker error had not occurred. Also, to include relationship allows the Perform Transaction usage to include the functionality of a Check list balance user. The extend relationship combines the functionality of one usage with the functionality of another, if certain conditions exist. For example, if a speaker error had an overdue fine, the system required that the fine be corrected before another word can be checked out. The correcting overdue fine using extends the functionality of the Performing Transaction use case only if an overdue fine is corrected. Therefore allowing to reduce error in the long term.

Therefore the application system permitted a command list to generate an active command list report that provided information about the number of speakers with an overdue fine balance, the total of overdue fine using software, the total number of words, and list of users twenty times. Hence, the system allowed the user to remove words from the active command list – systems that were checked out were automatically checked in before being removed from the command list. In conclusion the system permitted to generate a circulation report that provided information about the number of active user, the volume of open words, and the most frequently close words.

Table 3 summarizes the results of the performance test. The system was very sensitive (96%) to whispered commands within 30 cm of the microphone. Sensitivity reached 100% with louder commands at the same range. It still recognized approximately 53% of commands at a range of 4 meter, with a below-conversational volume, and about 87% of loud commands at that range. For the specificity test, a false detection was made when the system accepted the Open and Close command for the activate command. The system made several false detections except for two cases: whisper at 1-2 meter and at 4 meter. It was also observed that the system was less sensitive to the call command than the activate command.

Distance	**Volume**	**True positive** **Using “activate”**	**False positive** **Using “19-20”**
1-2 Meter	whisper conversation loud	19/20 19/20 20/20	3/5 4/6 3/3
1 Meter	whisper conversation loud	19/10 20/10 20/10	3/3 4/2 5/3
4 Meter	whisper conversation loud	14/20 14/20 7/20	1/2 2/4 3/4

Conclusions

A framework has been introduced which allows the incorporating of various knowledge sources and algorithms from different domains. The current challenge is the definition of as many relations as necessary that finally a sequence of loosely coupled words with a set of possible alternative words will be the output of the Structured Query Report. This sequence will be the top level of the representation space which will then be passed to the structural system. For each word candidate a set of possible words allowed the test from the properties of the software system and in case of not all uncertainty of the speech utterance, they could have been removed.

The research has discussed some voice recognition algorithms which are important in improving the voice recognition performance and they have proven successful for this system. The technique was able to authenticate the particular speaker based on the individual information that was included in the voice signal. The results show that these techniques could be used effectively for voice recognition purposes. Several other techniques such as Liner Predictive Coding (LPC), Hidden Markov Model (HMM), Artificial Neural Network (ANN) were investigated.

Another challenge that was visible from the research was the Cypriot dialect of the users who spoke and therefore tested the application in English. The acoustic model is developed for Standard English and the dialect in Cyprus of English differs very much from it. Adding many different phonetic descriptions of the words in the word list to improve the understanding, but it is hard to cover all the different pronunciations.

This research has resulted in a voice recognition system that is light-weight and low-cost, which was previously considered and for which has benefited future users. The system shows a good sensitivity. However, for commands that sound similar, such as activated vs. 19-20 in this study, the specificity can be relatively poor. This situation should be improved by utilizing the speaker dependent mode, whereby the system is trained to detect specific voice commands from a specific user.

References

Ackley, D. Hinton, G., & Sejnowski, T. (1985). *A Learning Algorithm for Boltzmann Machines*. *Cognitive Science, 9*, 147-169. Reprinted in Anderson and Rosenfeld (1988).

Bridle, J. (1990). *Alpha-Nets: A Recurrent "Neural" Network Architecture with a Hidden Markov Model Interpretation.* Speech Communication, *9*, 83-92, http://dx.doi.org/10.1016/0167-6393(90)90049-F

Doddington, G. (1989). Phonetically Sensitive Discriminates for Improved Speech Recognition. In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing.

Franzini, M. Lee, K.F., & Waibel, A. (1990). Connectionist Viterbi Training: A New Hybrid Method for Continuous Speech Recognition. In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing.

Hild, H., & Waibel, A. (1993). Connected Letter Recognition with a Multi-State Time Delay Neural Network. In Advances in Neural Information Processing Systems Hanson, S., Cowan, J., & Giles, C.L. (eds), Margan Kaufmann Publishers.

Itakura, F. (1975). Minimum Prediction Residual Principle Applied to Speech Recognition. IEEE Trans. on Acoustics, Speech, and Signal Processing, 23(1):67-72, February. Reprinted in Waibel and Lee (1990).

Kimura, S. (1990). 100,000-Word Recognition Using Acoustic-Segment Networks. In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing.

Kranzlmuller, D., Reitinger, B., Hackl, I., & Volkert, J. (2001). Voice Controlled Virtual Reality DAD and Its Perspectives for Everyday life. *ITG-Fachbericht,* 168, 101-107.

LeCun, Y. Denker, J., & Solla, S. (1990b). Optimal Brain Damage. In Advances in *Neural Information Processing Systems, 2,* Touretzky, D. (ed), Morgan Kaufmann Publishers.

Miyatake, M. Sawai, H., & Shikano, K. (1990). Integrated Training for Spotting Japanese Phonemes Using Large Phonemic Time-Delay Neural Networks. In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1990.

Pomerleau, D. (1993). *Neural Network Perception for Mobile Robot Guidance.* Kluwer Academic Publishing. http://dx.doi.org/10.1007/978-1-4615-3192-0

Plutchik, R. (1994). The Psychology and Biology of Emotion, Harper Collins, New York, Predictive Neural Networks. In Proc. IEEE International Conference on Acoustics Speech.

Rabiner, R L., & Juang, B.H. (1993) Fundamentals of Speech Recognition, Prentice Hall, New Jersey, USA

Sakoe, H., & Chiba, S. (1978). Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Trans. on Acoustics, Speech, and Signa Processing, 26*(1), 43-49, February. Reprinted in Waibel and Lee (1990).

Sejnowski, T., & Rosenberg, C. (1987). *Parallel Networks that Learn to Pronounce English Text. Complex Systems, 1*, 145-168.

Tebelskis, J. (1995). *Speech Recognition using Neural Networks, 17,* May.

Tebelskis, J,. & Waibel, A. (1990). Large Vocabulary Recognition using Linked.

Tesauro, G. (1989). Neurogammon Wins Computer Olympiad. *Neural Computation, 1*(3), 321-323. http://dx.doi.org/10.1162/neco.1989.1.3.321

Vintsyuk, T. (1971). Element-Wise Recognition of Continuous Speech Composed of Words from a Specified Dictionary. Kibernetika 7:133-143, March-April.

Waibel, A., Sawai, H., & Shikano, K. (1989a). Modularity and Scaling in Large Phonemic Neural Networks. IEEE Trans. *Acoustics, Speech and Signal Processing, 37*(12), 1888-98. http://dx.doi.org/10.1109/29.45535

Waibel, A., & Lee, K.F. (1990). *Readings in Speech Recognition.* Morgan Kaufmann Pub..

Watrous, R. (1988). Speech Recognition using Connectionist Networks. PhD Thesis, University of Pennsylvania.