Evaluation of NETtalk as a means to extract phonetic features from text for synchronization with speech.

(HS-IDA-EA-98-119)

Christer Levefelt (a94chrle@ida.his.se)

Institutionen för datavetenskap
Högskolan i Skövde, Box 408
S-54128 Skövde, SWEDEN

Final year project at the Computer Science study programme during spring term 1998.

Instructor: Mikael Bodén

Evaluation of NETtalk as a means to extract phonetic features from text for synchronization with speech.

Submitted by Christer Levefelt to Högskolan Skövde as a dissertation for the degree of BSc, in the Department of Computer Science.

1998-09-21

I certify that all material in this dissertation which is not my own work has been identified and that no material is included for which a degree has previously been conferred on me.

Signed: _____

Evaluation of NETtalk as a means to extract phonetic features from text for synchronization with speech.

Christer Levefelt (a94chrle@ida.his.se)

Key words: NETtalk, phonetic features, artificial neural networks

Abstract

The background for this project is a wish to automate synchronization of text and speech. The idea is to present speech through speakers synchronized word-for-word with text appearing on a monitor.

The solution decided upon is to use artificial neural networks, ANNs, to convert both text and speech into streams made up of sets of phonetic features and then matching these two streams against each other. Several text-to-feature ANN designs based on the NETtalk system are implemented and evaluated. The extraction of phonetic features from speech and the synchronization itself are not implemented, but some assessments are made regarding their possible performances. The performance of a finished system is not possible to determine, but a NETtalk-based ANN is believed to be suitable for such a system using phonetic features for synchronization.

Contents

1 Introduction

Synchronization of text and speech is essential in many applications such as subtitling of films and TV programs. Naturally, it is possible to do this synchronization manually, but it becomes time-consuming for larger amounts of data and need for finer synchronization.

In cases where there exists a body of text and a recording of a person reading that text, with a one-to-one relationship between text and speech, it is theoretically possible to automate the synchronization process of the two. This dissertation examines the possibility of doing so by converting both text and speech to sets of phonetic features, which are then matched against each other to achieve synchronization.

1.1 Digital talking books with synchronized text

This dissertation focuses on one recent application for synchronization of text and speech: digital talking books. A talking book is a recording of a narrator's voice, reading aloud from a book, newspaper etc.

Labyrinten Data AB in Falköping, Sweden has, under commission from the Swedish Library of Talking Books and Braille, developed the DAISY[1] Digital Talking Book System [Lab96]. This is a system intended for digitally stored talking books and offers many advantages over older talking books, recorded on tape. The most notable ones are:

- It allows for fast, arbitrary access to any part of the recording.
- It is less bulky. That which previously took up several tapes now fits easily onto one CD-ROM disc.
- It allows for other information, such as pictures or text, to be presented along with the recording at specific points in time.

The latter feature gives the system the potential to be used in many applications besides talking books. It is well suited for other information systems where digitized speech is the primary information medium. For example, there are plans on presenting the text of books, word for word, simultaneously with the recording. One application for this feature would be training for people suffering from dyslexia. However, manual synchronization of the text and speech of an entire book would be far too time-consuming to allow that method to be used more than occasionally. Therefore, some method to automate this synchronization is needed.

1.2 Previous knowledge and research in related areas

To the knowledge of the author, there has been no previous research on the specific topic of synchronization of text and speech, although research in both speech recognition and speech synthesis as well as a general knowledge of phonetics is relevant.

[1] Digital Audio Information System.

1.2.1 On phonetics

Speech is divided into phonemes, defined as the smallest element of speech sound that indicates a difference in meaning [Ele95]. The number of phonemes varies between different languages, but according to [Ele95], there are generally 20 to 40 phonemes in one language. A phoneme is denoted by a symbol surrounded by slashes, e.g. the first phoneme in the English word “cat” is /k/. The International Phonetic Alphabet (IPA) is a standard for representing phonemes with symbols, developed by the International Phonetic Association (also IPA). It contains many special symbols not found in the English alphabet. There are many different notation systems for representing phonemes using the standard ASCII character set to facilitate a computerized representation. See Table 1 for a definition of symbols used in this dissertation.

The three fundamental articulatory classifications of phonemes according to [Sch94] are whether they are voiced (if the vocal cords are vibrating), the place where the phoneme is produced, and the manner in which it is produced.

Vowel sounds are produced when air exhaled from the lungs causes a vibration of the vocal cords, and consequently, they are voiced. The type of vowel sound that is produced is determined by the position of the tongue in the mouth. /i/, for example, is a high, frontal vowel.

Consonant sounds are the result of turbulence as air passes through a constriction somewhere in the vocal apparatus. Some consonants are voiced, but others are not. Places of articulation for consonants include lips (labial) and teeth (dental). Some manners of articulation for consonants are:

- Plosives or stop consonants, where a pressure is built up behind a complete closure, and then released in an explosion of air.
- Fricatives, where air is forced past a constriction, creating friction, which causes a noisy vibration.
- Affricates, which are produced similarly to plosives, but the air is released more gradually, creating a plosive that transforms into a fricative.
- Nasals, where air is prevented from escaping through the mouth, but passes out through the nostrils, creating resonances in the nasal and oral cavity.

/p/, for example, is a labial plosive, since there is a complete closure at the lips.

Articulation of phonemes in normal speech is not discrete according to [Wat92]. One phoneme gradually transforms into another; a process that is called co-articulation. Articulation of a phoneme is also influenced by surrounding phonemes. This is called adaptation. In extreme cases, the influence may be so strong that a phoneme takes on the qualities of another phoneme. This is called assimilation. An example is the word “ink”, where the /n/ becomes a /G/ (as in “sing”) because of influence from the /k/.

Articulatory classifications of a phoneme such as those used above (e.g. voiced, labial and plosive) are called phonetic features. The sets of phonetic features associated with different phonemes in this dissertation are defined in Table 1, copied from [SR86]. No definitions of the phonetic features used in [SR86] are given except for their names and the categories they belong to (see Table 3). Because of this, when discussing the

properties of phonetic features the standard phonetic definitions of these terms are assumed to apply.

1.2.2 On ANNs

[RN95] describes Artificial Neural Networks (ANNs) as a way of representing functions using networks of simple arithmetic computing elements modelled to behave in the same way as neurons, the cells in the brain. ANNs are usually structured in layers of nodes, with a layer of inputs, a layer of output nodes and optionally hidden layers in between (see Figure 1). There are connections between nodes in different layers, each connection having a weight assigned to it that determines the impact of that connection on the output of the network. An ANN is trained by feeding to the input nodes a set of input data for which the desired output data (target) is known. If the output of the ANN is equal to the target, nothing is done. If there is a difference between output and target, the weights of connections in the ANN are adjusted to reduce the difference. Using the back-propagation algorithm the weight changes are divided between contributing connections from previous layers according to the strength of the connection. If the ANN is designed properly, with training, it will perform better and better for the input data it has been trained with. One of the greatest advantages of ANNs, though, is their generalization of training data so that they are often able to produce an output that is nearly correct, even for input data they have never been presented with before. This makes them very useful for applications in pattern recognition, e.g. speech synthesis and speech recognition.

One problem is to determine the optimal design of an ANN (number of layers, number of nodes in each layer and how to place connections between nodes). A network too small will not be able to fully represent the correct function. A network too large will suffer from overfitting; i.e. it will memorize all training data but will not generalize the function very well. Also, the CPU-time for training raises rapidly with the number of connections in the ANN. Thus, the design of an ANN is a trade-off between different factors, and an adequate design is usually found by testing some different designs and picking the one that best suits your needs.

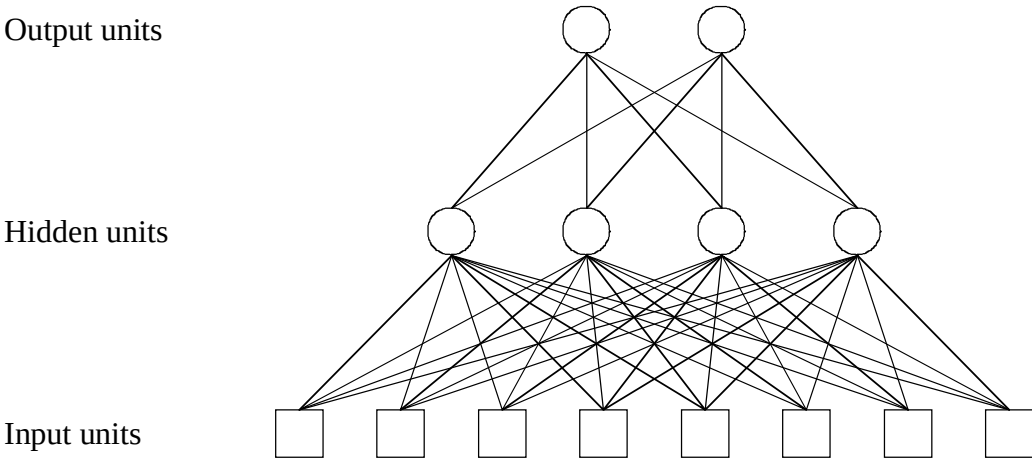

Figure 1: Sample ANN.

1.2.3 Speech synthesis using ANNs (The NETtalk System)

The NETtalk system [SR86], developed by T. J. Sejnowski and C. R. Rosenberg, is an example of a successful application of ANNs in the field of speech research. NETtalk is a system that is trained to convert English text into phonetic data and stress data that can be used in a speech synthesizer to produce synthetic speech. The system is built around an ANN that takes text as input and generates corresponding sets of phonetic features as output. Of course, this is not as simple as mapping one letter to a certain set of phonetic features. Instead, the system must take into consideration the context in which a letter occurs to determine the correct pronunciation. For example, in the sentence “This is a test.” the character sequence “is_” occurs twice, but in the first occurrence the “s” is unvoiced, /s/, but in the second it is voiced, /z/. To achieve this context sensitivity the text is fed to the ANN through a window, seven characters wide, which slides over the text. The central position of the window represents the letter whose corresponding phonetic features will be output from the ANN. The six letters in the positions preceding and following the central letter provide the context that is needed to determine what phonetic features should represent that letter. Thus, in the above example, the inputs and corresponding outputs for the first two occurrences of the letter “s” would be as shown in Figure 2.

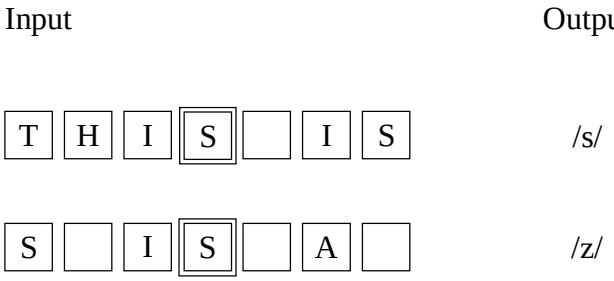

Figure 2: Sample input and output from the NETtalk system.

The letters have a local representation, i.e. each input unit encodes one character. There are 29 inputs, one for each letter of the English alphabet and three for punctuation characters. Several ANNs with a different number of hidden units (from 0 to 120) were tested [Sej88]. The phonemes have a distributed representation, i.e. all output units are used to encode one phoneme. There are 23 outputs, corresponding to different phonetic features, plus three outputs encoding stress and syllable boundaries. The phoneme set used in NETtalk together with some of the articulatory features of each phoneme is found in Table 1. The phoneme whose features most closely resemble the combination of output units is chosen as the output of the NETtalk system.

Two text sources were used for training:

- Continuous, informal speech of a child.
- A 20,012 word corpus from a dictionary.

The results were measured in two ways: “perfect match”, when the value of each articulatory feature was closer than 0.1 to the correct value (0 or 1), and “best guess”, the phoneme with features most closely resembling that of the output. An ANN with 80 hidden units achieved 95% correct best guesses and 55% perfect matches after 50,000 words of training on the informal speech corpus. An ANN with 120 hidden

Symbol	Phoneme	Articulatory features
/a/	f*a*ther	Low, Tensed, Central2
/b/	*b*et	Voiced, Labial, Stop
/c/	b*ough*t	Unvoiced, Velar, Medium
/d/	*d*ebt	Voiced, Alveolar, Stop
/e/	b*a*ke	Medium, Tensed, Front2
/f/	*f*in	Unvoiced, Labial, Fricative
/g/	*g*uess	Voiced, Velar, Stop
/h/	*h*ead	Unvoiced, Glottal, Glide
/i/	P*e*te	High, Tensed, Front1
/k/	*K*en	Unvoiced, Velar, Stop
/l/	*l*et	Voiced, Dental, Liquid
/m/	*m*et	Voiced, Labial, Nasal
/n/	*n*et	Voiced, Alveolar, Nasal
/o/	b*oa*t	Medium, Tensed, Back2
/p/	*p*et	Unvoiced, Labial, Stop
/r/	*r*ed	Voiced, Palatal, Liquid
/s/	*s*it	Unvoiced, Alveolar, Fricative
/t/	*t*est	Unvoiced, Alveolar, Stop
/u/	l*u*te	High, Tensed, Back2
/v/	*v*est	Voiced, Labial, Fricative
/w/	*w*et	Voiced, Labial, Glide
/x/	*a*bout	Medium, Central2
/y/	*y*et	Voiced, Palatal, Glide
/z/	*z*oo	Voiced, Alveolar, Fricative
/A/	b*i*te	Medium, Tensed, Front2 + Central1
/C/	*ch*in	Unvoiced, Palatal, Affricative
/D/	*th*is	Voiced, Dental, Fricative
/E/	b*e*t	Medium, Front1 + Front2
/G/	sin*g*	Voiced, Velar, Nasal
/I/	b*i*t	High, Front1
/J/	*g*in	Voiced, Velar, Nasal
/K/	se*x*ual	Unvoiced, Palatal, Fricative + Velar, Affricative (Compound: [k] + [S])
/L/	bott*le*	Voiced, Alveolar, Liquid
/M/	abs*ym*	Voiced, Dental, Nasal
/N/	butto*n*	Voiced, Palatal, Nasal
/O/	b*oy*	Medium, Tensed, Central1 + Central2
/Q/	*qu*est	Voiced, Labial + Velar, Affricative, Stop
/R/	b*ir*d	Voiced, Velar, Liquid
/S/	*sh*in	Unvoiced, Palatal, Fricative
/T/	*th*in	Unvoiced, Dental, Fricative
/U/	b*oo*k	High, Back1
/W/	b*ou*t	High + Medium, Tensed, Central2 + Back1
/X/	e*xc*ess	Unvoiced, Affricative, Front2 + Central1
/Y/	c*u*te	High, Tensed, Front1 + Front2 + Central1
/Z/	lei*s*ure	Voiced, Palatal, Fricative
/@/	b*a*t	Low, Front2
/!/	Nazi	Unvoiced, Labial + Dental, Affricative (Compound; [t] + [s])
/#/	e*x*amine	Voiced, Palatal + Velar, Affricative (Compound: [g] + [z])
/*/	*o*ne	Voiced, Glide, Front1 + Low, Central1 (Compound: [w] + [^])
/:/	logi*c*	High, Front1 + Front2
/^/	b*u*t	Low, Central1
/-/	Continuation	Silent, Elide
/–/	Word Boundary	Pause, Elide
/./	Period	Pause, Full Stop

Table 1: Articulatory representation of phonemes and punctuation used in NETtalk. Copied from Table 1 in [SR86].

nodes was trained on the 1,000 most common English words and then trained with 5 passes through the dictionary. After this it achieved 90% correct best guesses and 48% perfect matches in the dictionary. The system is also trained to detect primary, secondary and tertiary stress as well as syllable boundaries.

The sets of articulatory features representing phonemes in the NETtalk system, described in Table 1, is not complete. Some of the features that make up a phoneme are not listed in the table. One example of this is vowels, which are listed without their inherent feature “Voiced”.

According to [SR86], outputs are represented in terms of 23 articulatory features and 3 additional features encoding stress and syllable boundaries. The following features are listed in Table 1 of [SR86]:

- Position in mouth: Labial/Front1, Dental/Front2, Alveolar/Central1, Palatal/Central2, Velar/Back1, Glottal/Back2
- Phoneme Type: Stop, Nasal, Fricative, Affricative, Glide, Liquid, Voiced, Unvoiced, Tensed
- Vowel Frequency: High, Medium, Low
- Punctuation: Silent, Elide, Pause, Full Stop

The Continuation phoneme, /-/, is a marker used in the NETtalk system to tell that a certain letter of a word does not represent a phoneme by itself, but is silent or part of a combination of letters representing a phoneme. For example, the word “aardvark” is represented in NETtalk by the phoneme sequence /a-rdvark/ (the second “a” is silent).

1.2.4 Speech recognition using ANNs

K. Elenius and G. Takács have experimented with speech recognition using ANNs. In [ET90], they present a system converting speech data (in Swedish and Hungarian) to phonemes. An interesting property of this system is that internally, phonetic features are extracted from the speech input by an ANN.

Three different speech materials were used, which were then labeled by a human phonetic expert:

- INTRED – Swedish sentences read in a natural way by a trained male speaker.
- JONSSON – Swedish sentences read by another male speaker. No results relevant to this dissertation are presented for this material.
- MAMO – Hungarian sentences read in a natural way by a male speaker.

Figure 3 shows the design of the [ET90] system. The sampled speech waveform is first fed to a filter bank with a Fast Fourier Transform procedure producing outputs in 16 frequency ranges. (The value of each output is affected by whether sound is within the appropriate frequency range.) These are used with a 10 ms interval as input to the “coarse neural net”, which is trained to recognize seven phonetic features: voiceness, noiseness, nasalness, frontness, centralness, backness and vowelness. The output of the ANN is then passed on through the “dual window data selector” to the “fine neural net”. The data from the feature network passses through a seven frames wide window with the three frames preceding and the three frames following the current frame. The output of the filter bank is also used directly, passing through a one frame

wide window. The outputs of the “fine neural net” are possible phoneme candidates. The node with the highest activation indicates the first phoneme candidate, the next to highest is the second candidate and so on. The “segmentation window data selector” uses the activation of the first phoneme candidates over a 15 frame wide window (150 ms) to pass on to the “segmentation neural net”. This has a single output unit, which is supposed to have a high activation only at the first frame of each phoneme. The “fine activity filter” smoothes the activation levels output from the phoneme network to avoid errors caused by short spurious activation peaks. The “phoneme candidate data selector”, finally, selects the most probable phoneme as the output of the system.

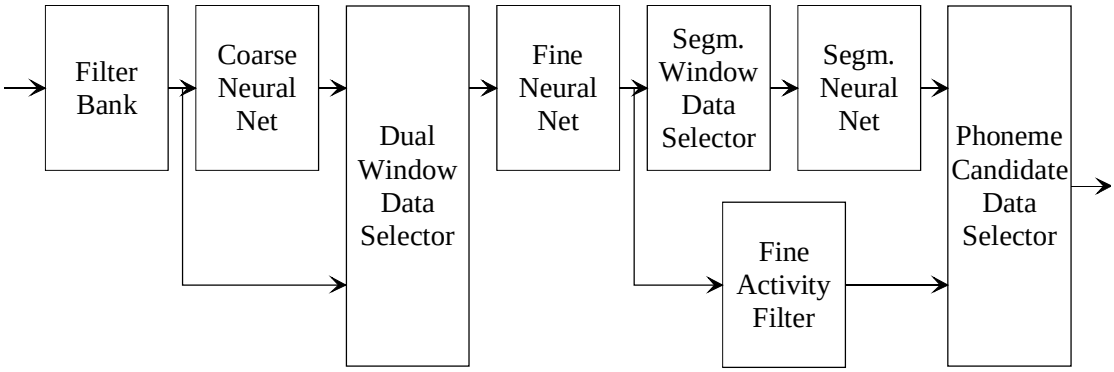

Figure 3: The architecture of the Elenius’ and Takács’ speech-to-phoneme system. Based on Figure 3 in [ET90].

Of the units in Figure 3, only the fine neural net and the phoneme candidate data selector contain language specific elements. To use the system for another language requires changing these units to accommodate to the new phoneme set, and to retrain the phoneme network.

Two variations of the system was tested, one as described above, and one where no coarse neural net was used and the only input to the fine neural net was the outputs of the filter bank. The system using the feature network showed considerably better performance. In this system, 62.4% of the correct phonemes were selected as the first phoneme candidate and 80.8% were among the first three candidates. The system only using inputs directly from the filter bank, the corresponding figures were 39.8% and 59.4%.

Table 2 shows the performance of the feature extraction ANN. The manner of articulation features performed generally better than the place of articulation features. Testing on the training set resulted in just 2% better results than the test set, indicating the network was good at generalization of data. The system was also tested for speaker and language independence, and the recognition rate of the features changed only a little compared to results from the same speech material. This shows that the features used are quite robust to change of speaker and even change of language.

Feature	Correct feature recognition (%)	
	INTRED	**MAMO**
Voiceness	93.1	93.3
Noiseness	91.0	92.9
Nasalness	95.4	93.1
Frontness	81.7	88.4
Centralness	83.2	80.8
Backness	88.7	88.2
Vowelness	88.2	88.0
All features correct	76.9	80.0

Table 2: Performance of the coarse phonetic feature network on the frame level when evaluating the feature activations as binary signals. From Table VII in [ET90].

2 Problem Definition

The goal of this dissertation is to present an evaluation of the performance of a system for synchronization of text and speech using a NETtalk-based ANN to extract phonetic features from text. Ideally, this system should be able to synchronize the speech and corresponding text of any English-language Talking Book, without any human interaction.

These features are desired in the system:

- The system should be able to synchronize correctly at the word level.
- The system should have an unlimited vocabulary.
- The synchronization should be speaker-independent.
- The synchronization should be automatic.

The system is not supposed to operate in real-time, i.e. no time limits are imposed on the system. When the synchronization process is complete, it is supposed to be verified and if necessary corrected by a human.

The following assumptions are made concerning the system and its input data:

- The speech is assumed to be in the form of a digitized recording and the text is also assumed to be stored in digital form. Both should be available for random access from some permanent storage device.
- There has to be a one-to-one relationship between the text and the speech data, i.e. the narrator is reading the text exactly as it appears in the text data. This is not always the case, e.g. abbreviations would present trouble.

2.1 The level at which synchronization should take place

The synchronization of the system is supposed to take place at the word level. Therefore any internal synchronization must take place at the word level or below by inserting synchronization points at specific points in time between some units recognizable from both the text and the speech data

Possible units for synchronization are:

- Words – Word boundaries are obviously easily detected in text, they occur at white space characters. Detection of word boundaries in speech, on the other hand, is very difficult, since when speaking normally, we rarely pause between words, but instead the stream of phonemes over several words is continuous.
- Letters – Letters are also easily detected in text, but unfortunately there is no such thing as letters in speech.
- Phonemes – To a human, phonemes are easily detected in speech, but to a computer it could prove difficult to analyze the data of digitized speech. With the pattern recognition capabilities of ANNs, however, it should be possible to achieve a fairly accurate detection of phonemes in speech, as indicated in [ET90]. Also, as described above, the NETtalk system is an example of a successful conversion from words to phonemes.

With this in mind, the obvious choice is to synchronize on the phoneme level in the internal representation of the system.

2.2 Design of a synchronization system

The synchronization system can be divided into three distinct modules (see Figure 4). One module takes a stream of text data as input, converting it to a stream of phonemes marked with the placement of word boundaries. Another module takes a stream of speech data as input, converting it to another stream of phonemes, this one encoded together with a timestamp for each phoneme. The third module attempts to synchronize the two phoneme data streams by matching the two phoneme streams and combining the information of the time at which phonemes occur with that of between which phonemes word boundaries occur to produce the information necessary for the synchronization.

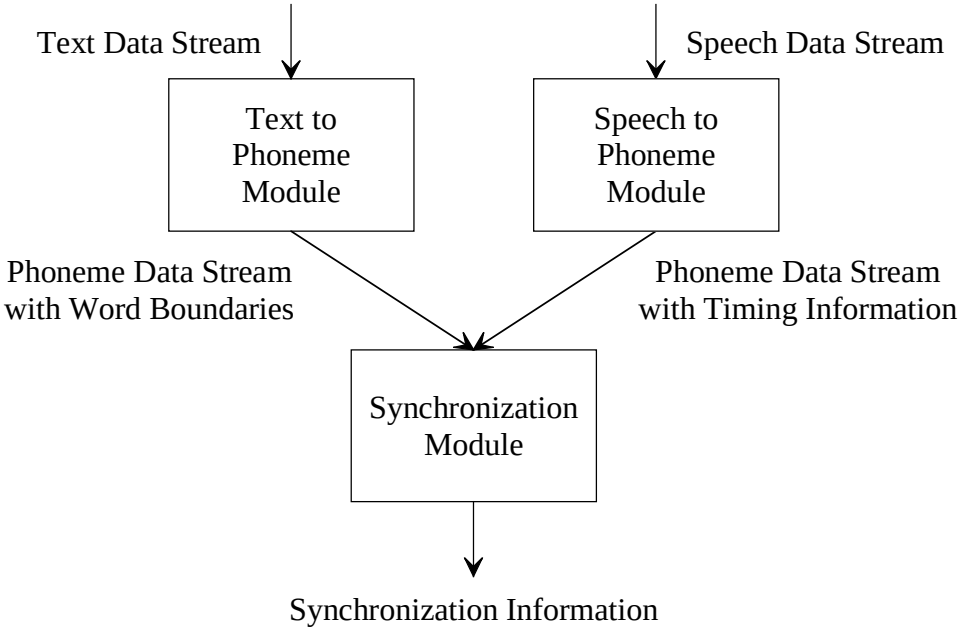

Figure 4: A text-speech synchronization system divided into modules.

2.3 Representation of phonemes

The two systems examined above, [SR86] and [ET90], both use a distributed representation of phonemes where they are represented by a set of phonetic features. [ET90] also observes this to be quite robust representation.

For this particular application, using phonetic features should provide a more stable synchronization. When converting text/speech data to phonemes, the output is either correct or incorrect. If a phoneme is incorrect, it is impossible to match it against the corresponding phoneme from the other data stream. With the distributed representation of the features, there is a large set of units to match. This should provide a greater tolerance against errors, since the system would search for the closest match between the feature sets. Even if some features are incorrect, the feature set could still be matched correctly against the other data stream.

2.4 Alternative representations of data streams

The decision to use a NETtalk-based system for this dissertation entails using phonemes as the internal representation of text and speech data. Another approach would be to convert one of the original representations to the other before synchronization.

- Conversion of speech to text

  This would require speaker-independent speech recognition for continuous speech with an unlimited vocabulary and a high recognition rate. Such a system is not feasible today, and will probably not be in the foreseeable future. Therefore, plain text is apparently not suitable as a representation for synchronization.

- Conversion of text to speech

  This would require a speech synthesizer for the data conversion. The greatest difficulty would then be to match the speed of the artificial speech with that of the sampled one, but this is not impossible. This approach would be interesting to examine, but will not be further discussed in this dissertation.

2.5 Evaluation of the synchronization system

A synchronization system such as the one outlined in section 2.2 is quite complex and the time constraints of the Final Year Project Course have not allowed for implementation of a complete system. Therefore, this dissertation will focus on the text-to-phoneme module of the system. Nevertheless an assessment is made regarding the feasibility and performance of the other two modules pictured in Figure 4. The three modules will be evaluated as indicated below:

- The Text-to-Phoneme Module. This module must achieve a high recognition rate of phonemes in the text data. This rate will be examined by the training and testing of a series of ANNs modeled after NETtalk, but modified for this application.
- The Speech-to-Phoneme Module. This module must achieve a high recognition rate of phonemes in the speech data. This rate will be assessed by examination of the results and speculations presented in [ET90].
- The Synchronization Module. This module must be able to combine the output of the previous two modules to produce correct synchronization information for as many words as possible. This will be assessed by examination of the expected recognition rates of the two previous modules.

3 Implementation

A series of ANNs was implemented, modeled after the NETtalk system. Their goal was to output the correct set of features associated with the phoneme corresponding to the central character of a seven-character input window sliding over the input text. The other six characters provided the context necessary to make a correct decision (as described in section 1.2.3). The ANNs were implemented in C++ and used a C++ class with operations for ANNs, based on C functions developed by Mikael Bodén at the Department of Computer Science, University of Skövde.

The reason for implementing these ANNs was to get results on the performance of NETtalk-style ANNs in detecting phonetic features in text. [SR86] only presents results regarding the performance of the entire NETtalk system (including the decision of which phoneme best matches the detected phonemes).

After training and testing of the ANNs, the discovery was made that the feature sets used, taken from [SR86], are not complete (according to the standard definition of the phonetic terms). Some features are not listed with all phonemes they belong to; e.g. many vowels are listed without the “Voiced” feature, which is inherent in all vowels. Also, the “Unvoiced” feature should not be needed since “Voiced” and “Unvoiced” are complementary properties. This has probably affected the performance of the ANNs. With a more accurate and consistent representation of which features belong to certain phonemes, the ANN should be able to achieve a better generalization of the connection between input and output data.

3.1 Design of the ANNs

Nine ANNs were trained and tested, three each of three different designs. The designs differed only in the number of hidden nodes. The first had no hidden layer (see Figure 5), the second had one hidden layer of 80 nodes (see Figure 6) and the third one hidden layer of 120 nodes (see Figure 7). The number of hidden nodes in the different designs was chosen according to those used in the NETtalk ANNs [SR86]. The design with no hidden layer was included to get a measure of the effectiveness of using hidden nodes. Each layer was fully connected to the immediately preceding and following layers, i.e. each node in each layer was connected to all nodes of the preceding and following layers.

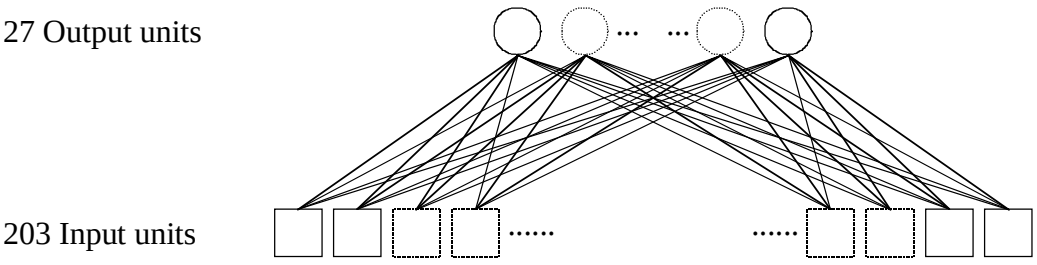

Figure 5: The ANN design with no hidden layer.

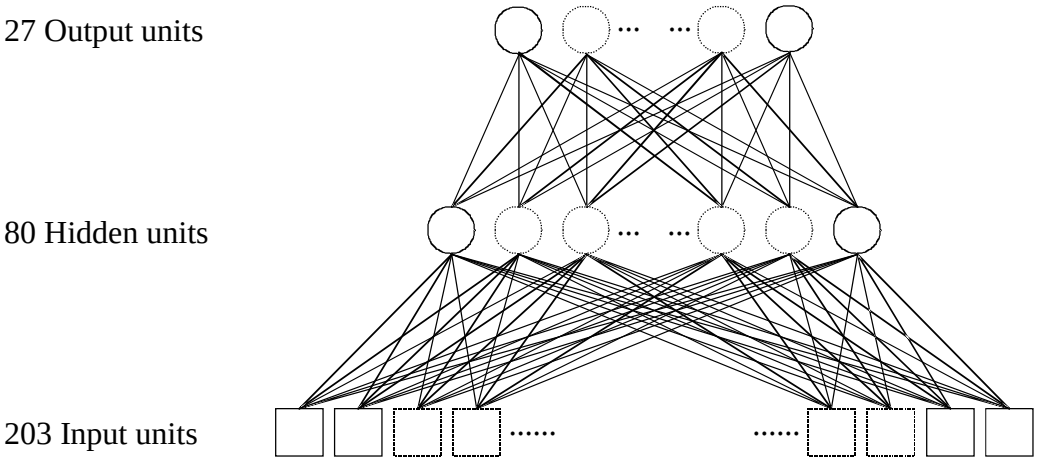

Figure 6: The ANN design with one hidden layer of 80 nodes.

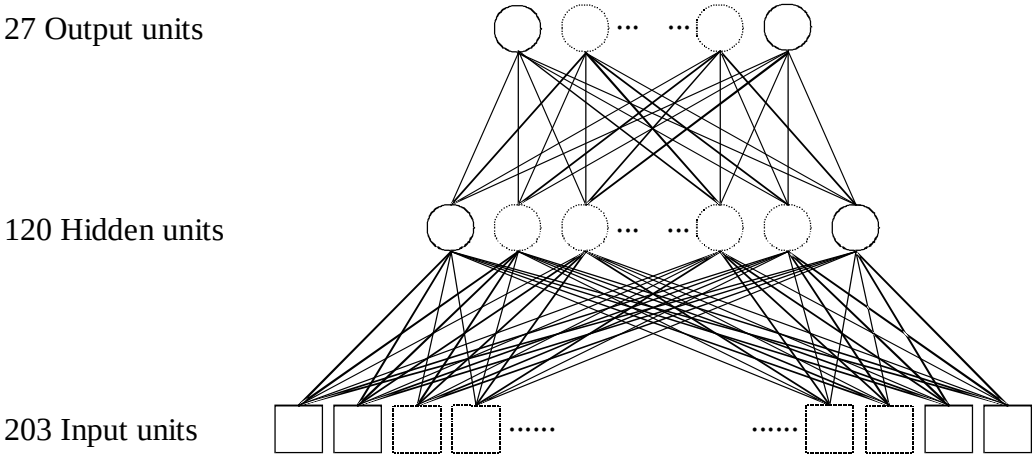

Figure 7: The ANN design with one hidden layer of 120 nodes.

The input layer of each network consisted of 203 input units in seven groups of 29 units, each group encoding one character, the seven groups forming the input window. The 29 units corresponded to the letters A-Z, a blank space and to the punctuation marks "." and ",". Any other characters were ignored by the system. The output layer consisted of 27 units, each corresponding to one phonetic, punctuation, stress or syllabic feature (see Table 3). The set of features used as output is based that used in the NETtalk system, but since [SR86] does not provide a complete list of which features are used, there may be some difference. The features in the first four categories of Table 3 (Position in mouth, Phoneme Type, Vowel Frequency and Punctuation) are listed with their respective phonemes in Table 1.

Category	**Features**
Position in mouth	Labial/Front1, Dental/Front2, Alveolar/Central1, Palatal/Central2, Velar/Back1, Glottal/Back2
Phoneme Type	Stop, Nasal, Fricative, Affricative, Glide, Liquid, Voiced, Unvoiced, Tensed
Vowel Frequency	High, Medium, Low
Punctuation	Silent, Elide, Pause, Full Stop
Stress	Unstressed, Primary Stress, Secondary Stress
Syllable Information	Start of Syllable, End of Syllable

Table 3: Phonetic features and other features used as outputs of the ANNs.

3.2 Training of the ANNs

The training and test data was taken from [Sej88][2], a data corpus consisting of about 20,000 words with four fields of information for each word (see Table 4):

- A letter representation.
- A phonetic representation.

  This used the same phoneme set defined in Table 1 with two exceptions. The phoneme represented by /:/ in Table 1 is not used in [Sej88], which instead includes /+/, described as the diphthong "oi" in French loan-words such as "abattoir" and "mademoiselle".

- Stress and syllabic structure.

  ">" indicates a consonant prior to a syllable nucleus.

  "<" indicates a consonant or vowel following the first vowel of a syllable nucleus.

  "0" indicates the first vowel in the nucleus of an unstressed syllable.

  "1" indicates the first vowel in the nucleus of a syllable receiving primary stress.

  "2" indicates the first vowel in the nucleus of a syllable receiving secondary stress.

- An indicator of foreign and irregular words.

  "1" indicates an irregular word.

  "2" indicates a foreign word.

  "0" is used for all other words.

The few (21) words that included the /+/ phoneme were removed from the corpus before training since no description of the phonetic features comprising it was available. "Phonemes" listed in Table 1 which in fact are compounded of two consecutive phonemes, e.g. /K/, were assigned all the features of the two phonemes.

Plain text	Phonetic transcription	Stress and syllables	Foreign word
aardvark	a-rdvark	1<<<>2<<	0
aback	xb@k-	0>1<<	0
abacus	@bxkxs	1<0>0<	0
abaft	xb@ft	0>1<<	0
abalone	@bxloni	2<0>1>0	0
abandon	xb@ndxn	0>1<>0<	0
abase	xbes-	0>1<<	0
abash	xb@S-	0>1<<	0
abate	xbet-	0>1<<	0
abatis	@bxti-	1<0>2<	2

Table 4: Example from the [Sej88] data corpus.

[2] The [Sej88] corpus may be used free-of-charge for non-commercial research purposes. [Sej88] contains details on whom to contact for more information on commercial use.

The data was randomly divided into two sets, one to use for training (approximately 80% of the data) and one for testing (the remaining 20%). The training set consisted of 17,953 words with a total of 150,306 characters. The test set consisted of 2,055 words with a total of 17,257 characters. After this, the data sets were converted to binary data for training and testing of the ANNs (see Figure 8). In this conversion, the irregular/foreign word indicator was ignored, which might have negatively affected the performance of the ANNs.

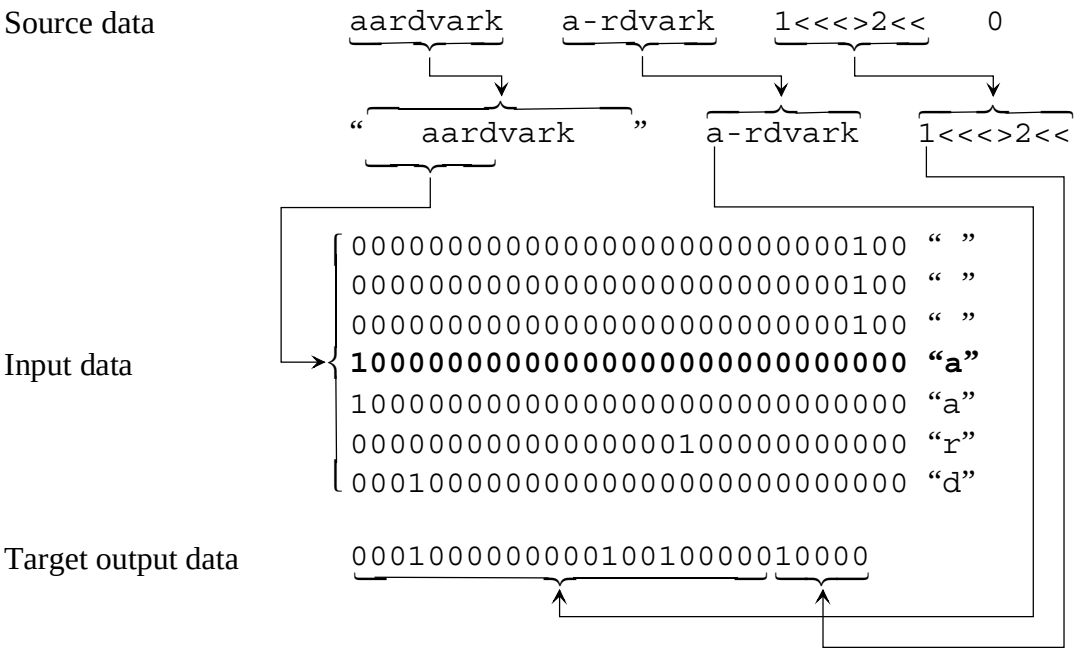

Figure 8: Conversion of the first letter of the word “aardvark” from the [Sej88] data to training data.

Each word of the training and testing sets were fed separately through the ANNs. This was achieved by inserting three white space characters before and after every word so that no two words would be inside the seven-character window at the same time. The reason for this was to prevent training of one word from being influenced by the preceding and following words. When training the ANN using complete sentences this influence should be exploited (teaching the ANN any relations that might exist between the ending of one word and the beginning of the next word). Since the data consisted of separate words as opposed to sentences, however, this would likely teach the ANN incorrect relations, if any at all.

The words of the training data were fed to the ANNs in alphabetical order. A random order of processing could possibly have generated better results by eliminating the risk for the ANNs to learn incorrect rules such as “All words begin with an A” and then having to unlearn this when reaching words starting with B.

Each ANN underwent ten epochs of training. After every epoch, the ANN was tested using the test set. A learning rate of 0.9 and a momentum term of 0.2 were used throughout the training, values corresponding to those used in the original NETtalk system [SR86].

3.3 Performance of the ANNs

All training and testing took place on a test machine with a 120 MHz Pentium CPU. The ANN with no hidden layer completed one epoch of training in approximately 20 minutes on this machine. The corresponding times for the ANNs with 80 and 120 hidden nodes were 81 and 121 minutes, respectively.

All performance statistics are based on the test set, independent from the training set. The symbols representing different features in Figure 9 to Figure 11 are presented in Table 5. The results shown in these figures, however, are not intended to show the performance of individual features, but to give a general idea of the different ANNs' relative performance and the improvement of performance with training. See Appendix A for complete test results of two of the ANNs, the best performing and the worst performing.

3.3.1 The design with no hidden nodes

These ANNs showed errors ranging from 3.5% to 5.5% after ten epochs of training. The average error was 4.1%. Most features had stabilized around a certain error by the end of the fourth epoch. Most features had around 4% errors after training was complete, but some features remained between 5% and 5.5% throughout the training (see Figure 9).

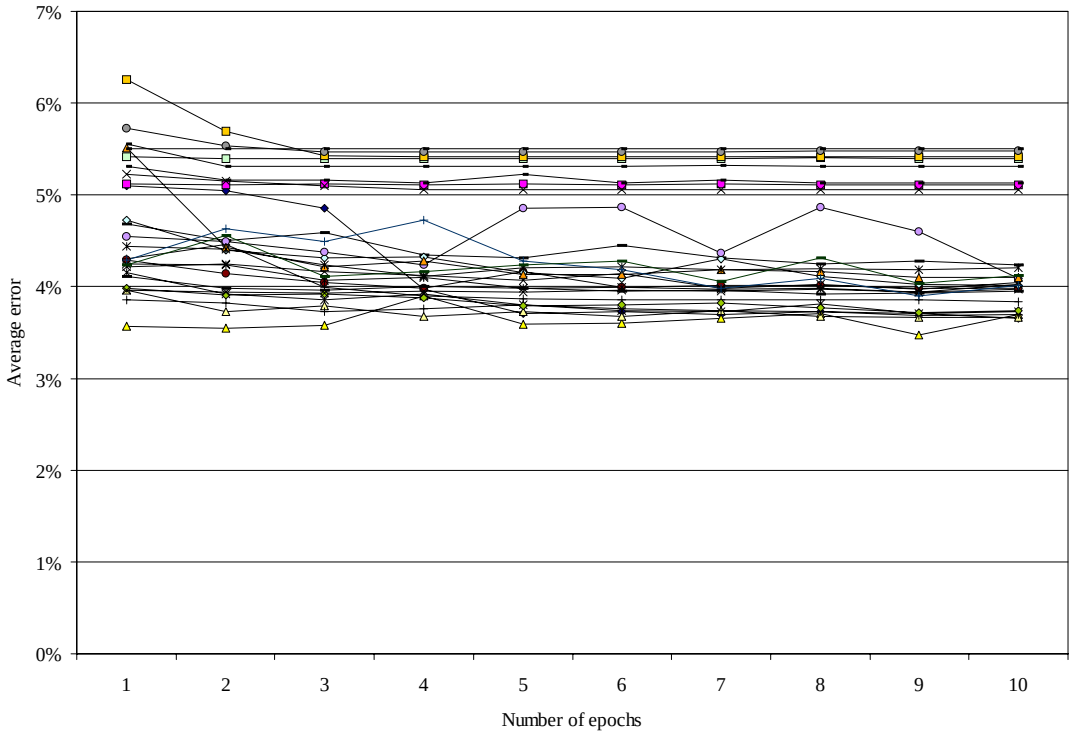

Figure 9: Typical test results for the ANNs with no hidden layer.

3.3.2 The design with 80 hidden nodes

These ANNs showed errors ranging from 2.7% to 5.2% after ten epochs of training. The average error was 3.3%. Most features had stabilized around a certain error by the end of the fourth epoch, but some dramatic improvements occurred even in the ninth

and tenth epoch. Here too, however, some features showed no improvement after the first epoch, but stayed around 5% error (see Figure 10).

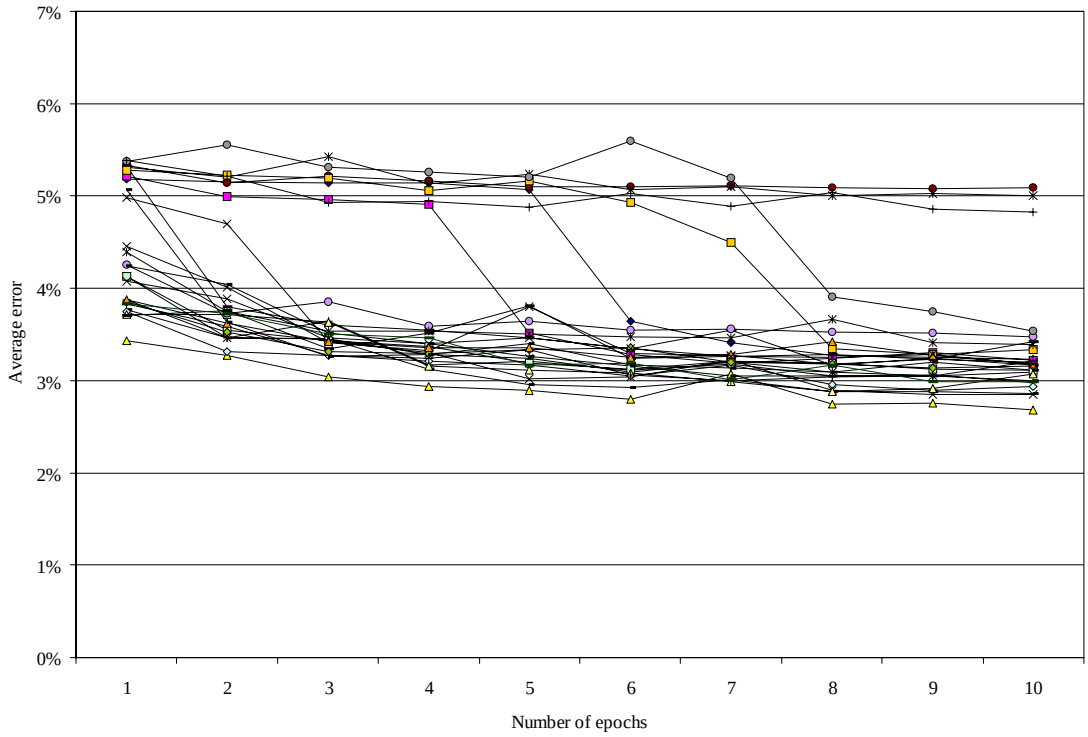

Figure 10: Typical test results for the ANNs with one hidden layer of 80 nodes.

3.3.3 The design with 120 hidden nodes

These ANNs showed errors ranging from 2.6% to 3.4% after ten epochs of training. The average error was 2.9%. No features showed any dramatic improvements after the fourth epoch, but small improvements were made up to the tenth and last epoch. In addition to showing better results overall, the performance of the features in this design was more consistent, no features showed significantly worse results from any others (see Figure 11).

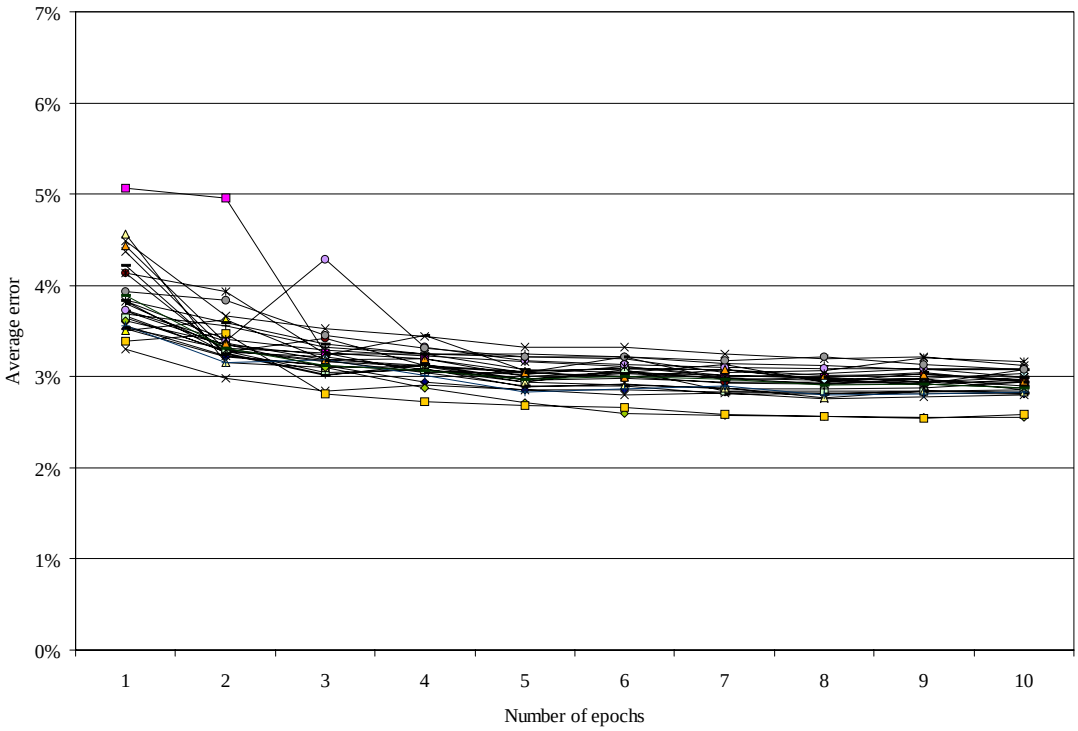

Figure 11: Typical test results for the ANNs with one hidden layer of 120 nodes.

3.3.4 Performance of individual features

Table 5 shows the performance of individual features in the 120-hidden nodes ANNs. As mentioned above, the performance of different features are very consistent. The difference between the highest and the lowest error is 0.8 percentage units.

Symbol	Feature	Minimum (%)	Maximum (%)	Average (%)
	Labial/Front1	97.1	97.3	97.2
	Dental/Front2	96.9	97.2	97.1
	Alveolar/Central1	97.1	97.4	97.3
	Palatal/Central2	97.1	97.3	97.2
	Velar/Back1	96.6	97.0	96.9
	Glottal/Back2	97.0	97.1	97.1
	Stop	96.9	97.3	97.0
	Nasal	96.9	97.2	97.1
	Fricative	96.9	97.1	97.0
	Affricative	97.0	97.2	97.1
	Glide	96.9	97.2	97.1
	Liquid	97.2	97.3	97.2
	Voiced	96.9	97.1	97.1
	Unvoiced	96.8	97.0	96.9
	Tensed	96.9	97.0	96.9
	High	97.1	97.3	97.2
	Medium	96.9	97.2	97.1
	Low	96.9	97.2	97.1
	Silent	97.1	97.4	97.3
	Elide	97.3	97.4	97.4
	Pause	97.1	97.1	97.1
	Full Stop	96.9	97.0	96.9
	Unstressed	97.0	97.1	97.1
	Primary	96.9	96.9	96.9
	Secondary	97.1	97.2	97.2
	Start of Syllable	97.0	97.1	97.1
	End of Syllable	96.7	97.2	97.0

Table 5: Performance of the different features for the ANNs with 120 hidden nodes. The symbols correspond to those used in Figure 9 to Figure 11.

3.3.5 Reliability of decisions

The data output from testing of the ANNs was divided into four categories according to whether it decided a feature was present or absent and whether it decided correctly or incorrectly. The categories were labeled according to Figure 12, i.e. C1 is the category of all outputs where the ANN correctly determined that the feature in question was present in the current phoneme. The division of results into these categories allows computation of some additional statistics, as shown below.

	Correct decision (C)	Incorrect decision (I)
The ANN decided the feature was absent in the feature (0)	C0	I0
The ANN decided the feature was present in the feature (1)	C1	I1

Figure 12: The different categories of test results.

Figure 9, Figure 10 and Figure 11 show typical testing results for the three different network designs after each epoch of training. The errors depicted represent all incorrect decisions made by the ANN, computed by the formula

$$\text{Error} = \frac{I0 + I1}{C0 + C1 + I0 + I1},$$

or put another way, the probability of making a correct decision is

$$\text{P(Correct guess)} = \frac{C0 + C1}{C0 + C1 + I0 + I1}.$$

Other statistics that are possible to compute are the probability of detecting the presence of a feature in a phoneme,

$$\text{P(Detect presence)} = \frac{C1}{C1 + I0},$$

the probability of detecting the absence of a feature in a phoneme,

$$\text{P(Detect absence)} = \frac{C0}{C0 + I1},$$

the probability of a decision being correct if feature is detected in a phoneme,

$$\text{P(Feature present | Feature detected)} = \frac{C1}{C1 + I1}$$

and the probability of a decision being correct if feature is not detected in a phoneme,

$$\text{P(Featue absent | Feature not detected)} = \frac{C0}{C0 + I0}.$$

Table 6 shows the average values of these probabilities for the ANNs with 120 hidden nodes. The low probability (around 50%) of detecting the presence of a feature compared to that of detecting the absence of one (over 99.5%) is probably because most features occurs only in a few phonemes (especially with the phoneme representation in Table 1). Therefore, the ANN has more training on detecting the absence of features than the presence thereof. The 120-node design, however,

performs better than the other designs, so the effect is not a result of overfitting (see section 1.2.2).

Features	Probability of making a correct decision (%)	Probability of detecting the presence of a feature (%)	Probability of detecting the absence of a feature (%)	Reliability of a decision given that the feature is detected (%)	Reliability of a decision given that the feature is not detected (%)
Labial/Front1	97.2	51.5	99.8	94.9	97.2
Dental/Front2	97.1	50.0	99.8	92.4	97.2
Alveolar/Central1	97.3	51.1	99.8	92.6	97.4
Palatal/Central2	97.2	52.5	99.7	89.4	97.5
Velar/Back1	96.9	50.2	99.6	86.6	97.2
Glottal/Back2	97.1	52.0	99.7	91.6	97.3
Stop	97.0	48.6	99.8	94.5	97.1
Nasal	97.1	47.7	99.8	92.6	97.2
Fricative	97.0	51.3	99.6	89.3	97.3
Affricative	97.1	48.1	99.8	92.7	97.2
Glide	97.1	50.7	99.8	92.2	97.2
Liquid	97.2	50.7	99.7	91.2	97.4
Voiced	97.1	49.7	99.8	92.6	97.2
Unvoiced	96.9	46.9	99.7	90.4	97.1
Tensed	96.9	50.1	99.7	89.6	97.2
High	97.2	53.4	99.7	89.8	97.5
Medium	97.1	50.8	99.7	90.7	97.3
Low	97.1	51.8	99.7	90.0	97.3
Silent	97.3	50.6	99.8	93.9	97.4
Elide	97.4	54.1	99.8	93.3	97.5
Pause	97.1	48.3	99.8	93.7	97.2
Full Stop	96.9	47.5	99.8	92.8	97.1
Unstressed	97.1	48.8	99.8	92.5	97.2
Primary	96.9	49.5	99.7	90.0	97.1
Secondary	97.2	47.3	99.9	96.1	97.2
Start of Syllable	97.1	50.6	99.8	92.6	97.2
End of Syllable	97.0	48.9	99.7	91.7	97.2

Table 6: Average reliability of the ANNs with 120 hidden nodes.

3.4 Frequency of phonetic features

For the synchronization system it is not enough for a feature to be easily recognized by the text-to-phoneme module and by the speech-to-phoneme module. If the synchronization module relies heavily on a specific feature or set of features, it must appear frequently enough in a text or synchronization will not be possible. Because of this, the frequency of phonetic features is an important measure of how useful they are for synchronization.

In order to get a notion of the frequency of different phonetic features a sample text of 2,308 words was converted semi-automatically into a phonetic representation. A simple program was then developed, which examined the phonetic data for the frequency of individual features. The results are shown in Table 7. As noted in section 1.2.1, the feature set used is not complete, which of course affects these figures. With a correct feature set, the results for some features would improve (significantly in the

case of “Voiced”). Since only phonetic data was examined, there is no data for stress and syllable features.

Feature	Distance (in characters)		
	Minimum	**Maximum**	**Average**
Labial/Front1	1	38	5
Dental/Front2	1	50	8
Alveolar/Central1	1	28	4
Palatal/Central2	1	60	8
Velar/Back1	1	96	12
Glottal/Back2	1	278	52
Stop	1	46	6
Nasal	1	96	12
Fricative	1	50	8
Affricative	4	846	172
Glide	3	666	100
Liquid	1	76	13
Voiced	1	20	3
Unvoiced	1	36	5
Tensed	1	86	11
High	1	61	10
Medium	1	54	7
Low	3	171	28
Silent	1	47	7
Elide	1	17	3
Pause	2	16	5
Full Stop	11	272	112

Table 7: Distance between individual features of the same type in the sample text.

4 Conclusions

For a synchronization system to be successful, some set of features must be found that can be extracted from both text and speech with small errors and occurs frequently enough (at least once in every word) so they can be matched in the synchronization module.

4.1 Feasibility and performance of a synchronization system

As stated in section 2.5, the evaluation of the system was divided into three parts: evaluation of the text-to-phoneme module, of the speech-to-phoneme module and of the synchronization module.

The performance of the entire system depends on that of the three modules. The performance of the synchronization module depends on that of the other two modules. An attempt has been made to assess the performance of the three modules and of the entire system, based on:

- The results of the text-to-feature ANNs presented in chapter 3.
- The results of the speech-to-feature ANN presented in [ET90].

4.1.1 The text-to-phoneme module

As shown in section 3.3, the network design with 120 nodes shows consistently better performance than the other two designs. This makes it is the obvious choice for use in the synchronization system.

The probability of making a correct decision ranged from 96.6% to 97.4%. The probability of detecting the presence of a feature ranged from just 43.1% to 56.0%, while the probability of detecting the absence of a feature was very high, between 99.4% and 99.9%. The reliability of decisions was quite good; between 82.6% and 97.5% when features are detected and between 96.9% and 97.6% when they are not detected.

This means that even though the ANN may fail to detect more than half of the features in a phoneme, the reliability of the detection of those features is on average more than 90%. A more complete phoneme set where some or all phonemes have more features assigned to them would probably result in detection of more features.

As a measure of what performance could be expected when combining all features, the recognition rates of phonemes in the original NETtalk system can be examined (see section 1.2.3).

Even though it is impossible to prove without a working implementation of a complete system, these results are probably enough to establish a more or less working synchronization given that the speech-to-phoneme module produces similar results. The use of a more accurate representation of phonemes should also increase performance somewhat in general and even more in some specific aspects.

In any case, an ANN based on NETtalk performs quite well, and although some refinement might increase its performance further, the results presented in this dissertation makes it already seem very suitable for this application.

4.1.2 The speech-to-phoneme module

Table 2 shows the results of the different features output from the feature recognition ANN used in [ET90]. Based on these results, an attempt has been made to evaluate how well a speech-to-phoneme-module using a similar ANN might perform.

The results are for Swedish (INTRED) and Hungarian (MAMO) data corpora, but corresponding results for English should not vary much from these, since Swedish and Hungarian show very similar results, despite Hungarian being more different from Swedish than English from Swedish.

Table 8 contains the features used in [ET90] and shows the corresponding features used in NETtalk. Noiseness and Vowelness features do not exist in [SR86]. The place-of-articulation features are not discrete and the correspondence of these features between the systems is only approximate.

Voiceness and Nasalness are the only features equivalent in both systems, but it should be possible to train a speech-to-feature ANN to recognize other features with results comparable to those presented in Table 2 (80.8% to 95.4%). [ET90] suggests that replacing place-of-articulation features with spectral features (such as “compact”, “diffuse” and “flat”) might improve performance.

Articulatory Features [ET90]	**NETtalk Phonetic Features [SR86]**
Voiceness	Voiced
Noiseness	N/A
Nasalness	Nasal
Frontness	Labial/Front1 & Dental/Front2
Centralness	Alveolar/Central1 & Palatal/Central2
Backness	Velar/Back1 & Glottal/Back2
Vowelness	N/A

Table 8: The features used in [ET90] and corresponding features in [SR86].

4.1.3 The synchronization module

The synchronization module takes as input the two data streams consisting of sets of phonetic features and tries to match the sets from one stream against those from the other. This requires that enough features have been detected in enough phonemes in both of the preceding modules so that some features are possible to be matched against each other.

The only two features in [SR86] and [ET90] that are directly comparable are Voiceness/Voiced and Nasalness/Nasal. They also those that are best recognized by the [ET90] ANN (between 93.1% and 95.4%). The 120-node NETtalk ANNs made on average 97.1% correct decisions for both of these features but detected their presence in just 47.7% and 49.7%, respectively. In the sample text, “Nasal” occurred at most at a distance of 96 characters, and at least 1 character. The average distance was 12 characters. This indicates that “Nasal” is not by itself a good feature to use for synchronization, but it could be used together with other features. The “Voiced” feature occurred most at a distance of 20 characters and at least 1 character. The average distance was 3 characters. This indicates “Voiced” may be a suitable feature for synchronization, especially if using a more accurate representation of phonemes,

in which case “Voiced” would be present in all vowel sounds and some consonant sounds.

It is very hard to evaluate the performance of a system that is not implemented and is also depending on two other systems, one of which is not yet implemented. However, provided that both of the other two modules could produce performances above 90% for several features they both have in common (something that definitely seems possible) a synchronization module should be able to match them against each other. If it will perform well enough to synchronize entire talking books without too much help from a human, though, is impossible to determine at this stage.

4.1.4 Speculation on the performance of a complete system

No performance requirements have been set for the synchronization to be successful, but a system performing worse than 95%, i.e. making any kind of error for more than 5% of the words would probably not be usable for any larger body of data.

If both the text-to-phoneme module and the speech-to-phoneme module would detect one feature at a rate of 90%, that feature would statistically be detected in both data streams at the same time at a rate of 81%. If several features showed similar results, this might be enough to achieve a 95% performance for the complete system. The synchronization module, however, would probably need some additional knowledge on speech, such as the normal length of phonemes.

4.2 Suggestions for further work

The obvious work that is missing in this dissertation is of course implementations of the speech-to-phoneme module and of the synchronization module. Besides these, the following are suggestions for changes and examinations that could to lead to increased performance of the text-to-phoneme module in a finished synchronization system:

- Use of a more accurate representation of phonemes and their respective features.

  The use of the phoneme representation from [SR86] has affected all test results from the text-to-feature ANNs and also the study of the frequency of phonetic features in a sample text. A more accurate representation would probably have made better test results possible.

- An examination of what features might be suitable for synchronization.

  A good representation of phonemes should also have phonetic features that are useful for the synchronization process. Such features could be determined by an examination of different features (of those used in this dissertation as well as in other sources). Factors that affect their usefulness are their frequency in text/speech and the ease with which they can be detected in both text and speech.

- Experimentation with different designs for text-to-feature ANNs.

  While the 120-node design ANN was the best of the three tested designs, it might be possible to come up with even better designs. A greater number of

hidden would probably increase performance, as long as it does not result in overfitting.

- Use of the irregular/foreign word indicator available in the [Sej88] corpus.
- Training of the text-to-feature ANNs in non-alphabetical order.
- Expansion of the sliding window.

  According to [Sej88], experiments with a window of 11 characters instead of 7 produced better results.

4.3 Final conclusions

It is likely that a synchronization system using a NETtalk-based text-to-phoneme ANN is possible to construct and get to function well enough to be useable, but with the lack of implementation of two thirds of the complete system, this is not possible to either prove or refute conclusively. However, in a system using phonetic features for synchronization, a NETtalk-based system is very likely to be suitable for extraction of these features from text.

References

[Ele95] Elert, C-C. (1995) *Allmän och svensk fonetik* sjunde upplagan, Norstedts Förlag AB, Stockholm

[ET90] Elenius, K. and Takács, G. (1990) *Acoustic-Phonetic Recognition of Continuous Speech by Artificial Neural Networks*, STL-QPSR No. 2-3/1990, Department of Speech Communication and Music Acoustics, KTH, Stockholm

[Lab96] Labyrinten Data AB (1996) *DAISY Digital Talking Book System Data Format Specification* Revision 1.0, Labyrinten Data AB, Box 132, 521 02 Falköping, Sweden

[RN95] Russell, S. and Norvig, P. (1995*) Artificial Intelligence: A Modern Approach*, Prentice-Hall, Inc.

[Sch94] Schmandt, C. (1994) *Voice Communication with Computers: Conversational Systems*, Van Nostrand Reinhold, New York

[Sej88] Sejnowski, T. J. (1988), *The NetTalk Corpus: Phonetic Transcription of 20,008 English Words*, currently available from the Department of Information and Computer Science at the University of California, Irvine at ftp://ftp.ics.uci.edu/pub/machine-learning-databases/undocumented/connectionist-bench/nettalk/

[SR86] Sejnowski, T. J. and Rosenberg, C. R. (1986) 'NETtalk: a parallel network that learns to read aloud' in *Neurocomputing*, Anderson and Rosenfeld (Eds.), MIT Press, 1988

[Wat92] Waters, G. (1992) 'Speech production and perception' in Speech processing, Rowden, C. (Ed.), McGraw-Hill International (UK) Limited, 1992

Appendix A: Selected Test Results

The following tables show the test results of two of the text-to-feature ANNs. The tables indicate for each feature and each epoch how the ANNs' decisions were distributed over the categories defined in Figure 12. Numbers represent how many decisions the ANN made per category per feature per epoch. Table 9 shows the results from the worst performing ANN, which had no hidden nodes. Table 10 shows the results from the best performing ANN, which had one hidden layer of 120 hidden nodes.

	Epoch 1				Epoch 2			
	C0	**C1**	**I0**	**I1**	**C0**	**C1**	**I0**	**I1**
Labial/Front1	16288	89	860	20	16293	94	855	15
Dental/Front2	16319	55	882	1	16320	55	882	0
Alveolar/Central1	16275	367	527	88	16250	395	499	113
Palatal/Central2	16324	31	866	36	16339	29	868	21
Velar/Back1	16173	368	567	149	16222	359	576	100
Glottal/Back2	16115	402	555	185	16145	398	559	155
Stop	16144	371	574	168	16063	424	521	249
Nasal	16329	11	893	24	16350	17	887	3
Fricative	16039	409	531	278	16082	398	542	235
Affricative	16067	375	531	284	16116	383	523	235
Glide	16313	9	931	4	16317	9	931	0
Liquid	16214	360	530	153	16245	369	521	122
Voiced	16212	318	617	110	16182	341	594	140
Unvoiced	16237	286	642	92	16182	343	585	147
Tensed	16090	382	568	217	16082	399	551	225
High	16205	387	526	139	16207	390	523	137
Medium	16279	19	916	43	16321	19	916	1
Low	16148	399	544	166	16176	393	550	138
Silent	16210	360	518	169	16208	374	504	171
Elide	16148	29	882	198	16251	23	888	95
Pause	16268	38	878	73	16158	335	581	183
Full Stop	16181	392	549	135	16177	400	541	139
Unstressed	16141	349	573	194	16141	356	566	194
Primary	16261	8	945	43	16293	9	944	11
Secondary	16146	371	527	213	16069	389	509	290
Start of Syllable	16304	3	947	3	16304	3	947	3
End of Syllable	16139	387	546	185	16062	409	524	262

	Epoch 3				Epoch 4			
	C0	**C1**	**I0**	**I1**	**C0**	**C1**	**I0**	**I1**
Labial/Front1	16276	143	806	32	16176	395	554	132
Dental/Front2	16318	55	882	2	16320	55	882	0
Alveolar/Central1	16246	394	500	117	16176	409	485	187
Palatal/Central2	16349	29	868	11	16357	28	869	3
Velar/Back1	16238	353	582	84	16224	356	579	98
Glottal/Back2	16162	396	561	138	16153	417	540	147
Stop	16162	406	539	150	16169	415	530	143
Nasal	16347	20	884	6	16346	26	878	7
Fricative	16062	404	536	255	16098	409	531	219
Affricative	16130	383	523	221	16123	388	518	228
Glide	16317	9	931	0	16317	9	931	0
Liquid	16228	374	516	139	16244	378	512	123

Appendix A: Selected Test Results

	Epoch 3				Epoch 4			
	C0	**C1**	**I0**	**I1**	**C0**	**C1**	**I0**	**I1**
Voiced	16179	360	575	143	16171	375	560	151
Unvoiced	16201	354	574	128	16192	357	571	137
Tensed	16101	400	550	206	16115	410	540	192
High	16231	383	530	113	16221	387	526	123
Medium	16321	19	916	1	16321	19	916	1
Low	16174	400	543	140	16168	398	545	146
Silent	16230	350	528	149	16215	373	505	164
Elide	16305	15	896	41	16307	15	896	39
Pause	16184	346	570	157	16158	361	555	183
Full Stop	16183	396	545	133	16183	393	548	133
Unstressed	16181	345	577	154	16198	349	573	137
Primary	16304	9	944	0	16304	9	944	0
Secondary	16092	390	508	267	16049	392	506	310
Start of Syllable	16304	3	947	3	16304	3	947	3
End of Syllable	16136	412	521	188	16130	408	525	194
	Epoch 5				**Epoch 6**			
	C0	**C1**	**I0**	**I1**	**C0**	**C1**	**I0**	**I1**
Labial/Front1	16183	435	514	125	16201	413	536	107
Dental/Front2	16319	55	882	1	16320	55	882	0
Alveolar/Central1	16241	397	497	122	16229	406	488	134
Palatal/Central2	16357	28	869	3	16357	28	869	3
Velar/Back1	16253	347	588	69	16262	350	585	60
Glottal/Back2	16117	421	536	183	16167	401	556	133
Stop	16185	404	541	127	16191	400	545	121
Nasal	16327	28	876	26	16344	27	877	9
Fricative	16105	407	533	212	16077	412	528	240
Affricative	16143	397	509	208	16166	385	521	185
Glide	16316	9	931	1	16316	9	931	1
Liquid	16237	376	514	130	16251	372	518	116
Voiced	16170	385	550	152	16167	375	560	155
Unvoiced	16201	368	560	128	16200	367	561	129
Tensed	15995	425	525	312	15991	426	524	316
High	16227	373	540	117	16234	375	538	110
Medium	16321	19	916	1	16321	19	916	1
Low	16173	397	546	141	16179	397	546	135
Silent	16226	376	502	153	16225	376	502	154
Elide	16307	15	896	39	16307	15	896	39
Pause	16188	356	560	153	16172	372	544	169
Full Stop	16176	401	540	140	16181	392	549	135
Unstressed	16181	350	572	154	16180	350	572	155
Primary	16304	9	944	0	16304	9	944	0
Secondary	16126	392	506	233	16151	384	514	208
Start of Syllable	16304	3	947	3	16304	3	947	3
End of Syllable	16119	407	526	205	16125	394	539	199
	Epoch 7				**Epoch 8**			
	C0	**C1**	**I0**	**I1**	**C0**	**C1**	**I0**	**I1**
Labial/Front1	16198	422	527	110	16198	415	534	110
Dental/Front2	16319	55	883	0	16320	55	882	0
Alveolar/Central1	16248	378	517	114	16221	397	497	142
Palatal/Central2	16356	28	870	3	16357	28	869	3
Velar/Back1	16269	344	591	53	16248	351	584	74
Glottal/Back2	16177	388	570	122	16172	393	564	128

	Epoch 7				Epoch 8			
	C0	**C1**	**I0**	**I1**	**C0**	**C1**	**I0**	**I1**
Stop	16192	399	547	119	16189	402	543	123
Nasal	16339	28	877	13	16343	28	876	10
Fricative	16103	409	531	214	16118	406	534	199
Affricative	16116	398	508	235	16148	400	506	203
Glide	16315	10	931	1	16315	9	932	1
Liquid	16236	376	515	130	16248	374	516	119
Voiced	16195	374	561	127	16187	375	560	135
Unvoiced	16199	373	556	129	16207	362	566	122
Tensed	16080	423	528	226	15990	427	523	317
High	16227	384	530	116	16228	386	527	116
Medium	16320	19	916	2	16320	20	915	2
Low	16175	398	546	138	16187	393	550	127
Silent	16220	377	502	158	16229	377	501	150
Elide	16307	15	896	39	16307	15	896	39
Pause	16159	375	541	182	16165	373	543	176
Full Stop	16182	394	547	134	16183	388	553	133
Unstressed	16183	352	570	152	16180	353	569	155
Primary	16304	9	944	0	16303	9	944	1
Secondary	16184	385	513	175	16173	379	519	186
Start of Syllable	16304	3	947	3	16304	3	947	3
End of Syllable	16161	396	537	163	16108	404	529	216
	Epoch 9				**Epoch 10**			
	C0	**C1**	**I0**	**I1**	**C0**	**C1**	**I0**	**I1**
Labial/Front1	16206	411	538	102	16205	422	527	103
Dental/Front2	16320	55	882	0	16320	55	882	0
Alveolar/Central1	16267	391	503	96	16220	399	495	143
Palatal/Central2	16357	28	869	3	16357	28	869	3
Velar/Back1	16262	355	580	60	16263	350	585	59
Glottal/Back2	16175	394	563	125	16171	398	559	129
Stop	16190	402	543	122	16195	400	545	117
Nasal	16343	29	875	10	16344	28	876	9
Fricative	16112	407	533	205	16119	407	533	198
Affricative	16162	400	506	189	16170	395	511	181
Glide	16316	9	931	1	16316	10	930	1
Liquid	16246	378	512	121	16253	371	519	114
Voiced	16196	375	560	126	16183	375	560	139
Unvoiced	16211	366	562	118	16194	368	560	135
Tensed	16039	425	525	268	16136	415	535	171
High	16235	386	527	109	16235	385	528	109
Medium	16320	20	915	2	16320	20	915	2
Low	16182	397	546	132	16182	394	549	132
Silent	16242	374	504	137	16235	377	501	144
Elide	16307	15	896	39	16307	15	896	39
Pause	16182	368	548	159	16175	375	541	166
Full Stop	16181	396	545	135	16183	389	552	133
Unstressed	16182	352	570	153	16181	351	571	154
Primary	16303	9	944	1	16303	9	944	1
Secondary	16208	376	522	151	16178	387	511	181
Start of Syllable	16304	3	947	3	16304	3	947	3
End of Syllable	16158	402	531	166	16144	402	531	180

Table 9: Results from ANN with no hidden nodes, 1st run.

Appendix A: Selected Test Results

	Epoch 1				Epoch 2			
	C0	**C1**	**I0**	**I1**	**C0**	**C1**	**I0**	**I1**
Labial/Front1	16247	422	527	61	16267	456	493	41
Dental/Front2	16215	441	496	105	16249	464	473	71
Alveolar/Central1	16180	436	458	183	16323	428	466	40
Palatal/Central2	16220	390	507	140	16278	448	449	82
Velar/Back1	16221	385	550	101	16284	422	513	38
Glottal/Back2	16256	402	555	44	16267	436	521	33
Stop	16220	361	584	92	16237	405	540	75
Nasal	16211	392	512	142	16258	421	483	95
Fricative	16117	430	510	200	16205	459	481	112
Affricative	16211	419	487	140	16277	438	468	74
Glide	16183	407	533	134	16259	398	542	58
Liquid	16224	421	469	143	16292	432	458	75
Voiced	16229	371	564	93	16254	403	532	68
Unvoiced	16246	350	578	83	16180	398	530	149
Tensed	16122	430	520	185	16227	446	504	80
High	16105	445	468	239	16229	461	452	115
Medium	16207	397	538	115	16237	447	488	85
Low	16074	438	505	240	16222	467	476	92
Silent	16272	361	517	107	16284	417	461	95
Elide	16261	392	519	85	16287	433	478	59
Pause	16191	368	548	150	16259	420	496	82
Full Stop	16195	382	559	121	16220	419	522	96
Unstressed	16222	343	579	113	16245	434	488	90
Primary	16187	407	546	117	16232	438	515	72
Secondary	16242	378	520	117	16291	420	478	68
Start of Syllable	16192	429	521	115	16235	453	497	72
End of Syllable	16192	424	509	132	16277	477	456	47
	Epoch 3				**Epoch 4**			
	C0	**C1**	**I0**	**I1**	**C0**	**C1**	**I0**	**I1**
Labial/Front1	16248	483	466	60	16256	474	475	52
Dental/Front2	16233	476	461	87	16250	483	454	70
Alveolar/Central1	16244	465	429	119	16306	469	425	57
Palatal/Central2	16304	455	442	56	16299	460	437	61
Velar/Back1	16274	462	473	48	16238	492	443	84
Glottal/Back2	16274	458	499	26	16255	480	477	45
Stop	16260	423	522	52	16279	429	516	33
Nasal	16311	436	468	42	16263	454	450	90
Fricative	16169	479	461	148	16250	473	467	67
Affricative	16271	449	457	80	16304	449	457	47
Glide	16262	414	526	55	16277	424	516	40
Liquid	16286	440	450	81	16283	467	423	84
Voiced	16280	401	534	42	16295	411	524	27
Unvoiced	16250	433	495	79	16283	428	500	46
Tensed	16262	428	522	45	16269	431	519	38
High	16268	457	456	76	16305	467	446	39
Medium	16262	449	486	60	16240	480	455	82
Low	16265	451	492	49	16268	476	467	46
Silent	16322	410	468	57	16340	448	430	39
Elide	16266	477	434	80	16301	482	429	45
Pause	16262	412	504	79	16296	418	498	45
Full Stop	16201	451	490	115	16248	454	487	68
Unstressed	16261	439	483	74	16288	449	473	47

	Epoch 3				Epoch 4			
	C0	**C1**	**I0**	**I1**	**C0**	**C1**	**I0**	**I1**
Primary	16231	461	492	73	16231	458	495	73
Secondary	16283	428	470	76	16310	424	474	49
Start of Syllable	16214	464	486	93	16253	469	481	54
End of Syllable	16250	486	447	74	16267	482	451	57
	Epoch 5				**Epoch 6**			
	C0	**C1**	**I0**	**I1**	**C0**	**C1**	**I0**	**I1**
Labial/Front1	16268	488	461	40	16283	477	472	25
Dental/Front2	16249	487	450	71	16243	484	453	77
Alveolar/Central1	16319	463	431	44	16319	452	442	44
Palatal/Central2	16299	465	432	61	16296	466	431	64
Velar/Back1	16251	487	448	71	16233	496	439	89
Glottal/Back2	16264	487	470	36	16241	487	470	59
Stop	16296	416	529	16	16285	420	525	27
Nasal	16304	448	456	49	16324	438	466	29
Fricative	16253	487	453	64	16257	486	454	60
Affricative	16320	437	469	31	16295	448	458	56
Glide	16258	447	493	59	16260	444	496	57
Liquid	16284	457	433	83	16307	459	431	60
Voiced	16286	427	508	36	16270	445	490	52
Unvoiced	16279	435	493	50	16275	426	502	54
Tensed	16236	469	481	71	16247	464	486	60
High	16300	474	439	44	16291	471	442	53
Medium	16245	474	461	77	16261	471	464	61
Low	16258	476	467	56	16267	482	461	47
Silent	16313	463	415	66	16334	455	423	45
Elide	16304	486	425	42	16282	498	413	64
Pause	16309	437	479	32	16293	448	468	48
Full Stop	16263	452	489	53	16275	441	500	41
Unstressed	16298	437	485	37	16287	441	481	48
Primary	16250	446	507	54	16256	454	499	48
Secondary	16312	420	478	47	16302	430	468	57
Start of Syllable	16254	473	477	53	16252	481	469	55
End of Syllable	16277	485	448	47	16295	472	461	29
	Epoch 7				**Epoch 8**			
	C0	**C1**	**I0**	**I1**	**C0**	**C1**	**I0**	**I1**
Labial/Front1	16282	482	467	26	16289	472	477	19
Dental/Front2	16259	485	452	61	16265	485	452	55
Alveolar/Central1	16324	452	442	39	16328	457	437	35
Palatal/Central2	16294	465	432	66	16321	461	436	39
Velar/Back1	16244	492	443	78	16230	499	436	92
Glottal/Back2	16275	486	471	25	16263	503	454	37
Stop	16276	436	509	36	16279	434	511	33
Nasal	16319	438	466	34	16333	440	464	20
Fricative	16248	495	445	69	16242	506	434	75
Affricative	16318	439	467	33	16305	452	454	46
Glide	16245	460	480	72	16245	460	480	72
Liquid	16283	468	422	84	16313	463	427	54
Voiced	16271	445	490	51	16296	454	481	26
Unvoiced	16276	453	475	53	16295	445	483	34
Tensed	16268	450	500	39	16269	445	505	38
High	16281	482	431	63	16310	474	439	34
Medium	16256	483	452	66	16265	480	455	57

	Epoch 7				Epoch 8			
	C0	**C1**	**I0**	**I1**	**C0**	**C1**	**I0**	**I1**
Low	16227	513	430	87	16261	506	437	53
Silent	16321	465	413	58	16350	450	428	29
Elide	16304	495	416	42	16298	514	397	48
Pause	16303	448	468	38	16315	443	473	26
Full Stop	16268	446	495	48	16269	459	482	47
Unstressed	16303	433	489	32	16308	431	491	27
Primary	16246	464	489	58	16252	462	491	52
Secondary	16323	418	480	36	16326	440	458	33
Start of Syllable	16229	489	461	78	16239	477	473	68
End of Syllable	16268	482	451	56	16262	477	456	62
	Epoch 9				**Epoch 10**			
	C0	**C1**	**I0**	**I1**	**C0**	**C1**	**I0**	**I1**
Labial/Front1	16294	469	480	14	16294	461	488	14
Dental/Front2	16276	486	451	44	16279	492	445	41
Alveolar/Central1	16338	451	443	25	16331	452	442	32
Palatal/Central2	16310	475	422	50	16315	475	422	45
Velar/Back1	16269	499	436	53	16241	503	432	81
Glottal/Back2	16261	500	457	39	16263	499	458	37
Stop	16278	442	503	34	16285	438	507	27
Nasal	16324	449	455	29	16331	441	463	22
Fricative	16233	515	425	84	16239	514	426	78
Affricative	16314	449	457	37	16317	433	473	34
Glide	16265	465	475	52	16257	470	470	60
Liquid	16323	455	435	44	16321	464	426	46
Voiced	16300	455	480	22	16297	460	475	25
Unvoiced	16292	439	489	37	16290	449	479	39
Tensed	16253	467	483	54	16249	474	476	58
High	16310	482	431	34	16300	489	424	44
Medium	16263	483	452	59	16260	484	451	62
Low	16269	506	437	45	16264	507	436	50
Silent	16354	454	424	25	16359	449	429	20
Elide	16278	518	393	68	16296	510	401	50
Pause	16311	443	473	30	16315	444	472	26
Full Stop	16265	460	481	51	16275	456	485	41
Unstressed	16300	447	475	35	16315	448	474	20
Primary	16244	463	490	60	16255	466	487	49
Secondary	16331	445	453	28	16336	431	467	23
Start of Syllable	16258	482	468	49	16269	485	465	38
End of Syllable	16296	472	461	28	16295	474	459	29

Table 10: Results from ANN with 120 hidden nodes, 3rd run.