Mixed Distance Measures for Optimizing Concatenative Vocabularies for Speech Synthesis: A Thesis Proposal

CUCS-310-87

Nathaniel Polish
Columbia University
Computer Science Department
New York, New York

November 6, 1987

Abstract

Synthesized speech from text-to-speech systems is generally produced from the concatenation of small units of speech. The concatenation process can be complex, involving smoothing and context dependent adjustments to the speech. The overall quality of the speech produced will depend in large part on the quality of the elements used for concatenation. Selection and evaluation of these elements has been done entirely by hand. The proposed work addresses the process by which these concatenative elements are created from a natural voice and optimized. The optimization uses distance measures which exploit detailed information on the structure of the speech signals.

Contents

List of Figures

1 The Problem

The problem to be examined in this thesis is the creation, extraction and optimization of concatenative vocabularies for the generation of speech signals. The back-end or synthesizer section of text-to-speech systems builds speech signals by means of the concatenation of low-level voice elements [Polish 86]. The choice of the concatenation strategy as well as the choice of the actual concatenative elements are difficult to make. We present here a brief survey of concatenative synthesizers used in text-to-speech systems to clarify the problem and elucidate the state of the art in speech synthesis.

1.1 Concatenative strategy

We will consider the output of the first stage of a text-to-speech system to be a list of phonemes. These phonemes are derived from either a dictionary lookup or from letter-to-sound rules. Throughout this discussion we will consider that there exist 42 different phonemes and so we are dealing with a language of 42 symbols [Elovitz et al. 76]. There is nothing magic about the number 42. There are a variety of ways of counting phonemes and for English one way results in 42 phonemes. The phoneme stream disambiguates pronunciation of certain words. For example, "office" can be "off+ice" (long i) or "of+fice" or the correct one "off+ice" (short i). The string of phonemes must then be turned into sounds by means of a synthesizer system.

The simplest way to convert the stream of phonemes into sounds is to pre-record 42 pieces of speech. Each of the pieces contains a different phoneme as spoken by a particular person. If the phonemes are encoded in an appropriate way, then pitch and amplitude can be easily adjusted as needed at playback time. One such method is linear predictive coding (LPC). For a review of LPC and other speech coding techniques see [Rabiner and Schafer 78], [Parsons 86], or [Polish 86]. This is, with a few minor additions, the system used in the Votrax [1] text-to-speech system.

Once the list of phonemes is converted into a list of pieces of sound to be concatenated, we still need to apply some fairly complex modifications to the overall sound pattern. In fact, it is conceivable that the original idea of stringing together recordings of the 42 phonemes could be made to sound natural if we knew how to massage the final sound well enough. It turns out that we do not know nearly enough to make bad speech sound natural; but we can certainly apply a wide variety of smoothing techniques that can help.

The resulting speech is intelligible but sounds very poor. The fact that the speech is intelligible is reasonable considering the fact that phonemes are semantically defined. As [Flanagan 72] puts it: "the language must be constructed of basic linguistic units which have the property that if one replaces another in an utterance, the meaning is changed. This basic linguistic element is called a phoneme." In order to make the resulting speech more natural we need to consider some of the properties of the physical system that generates natural speech.

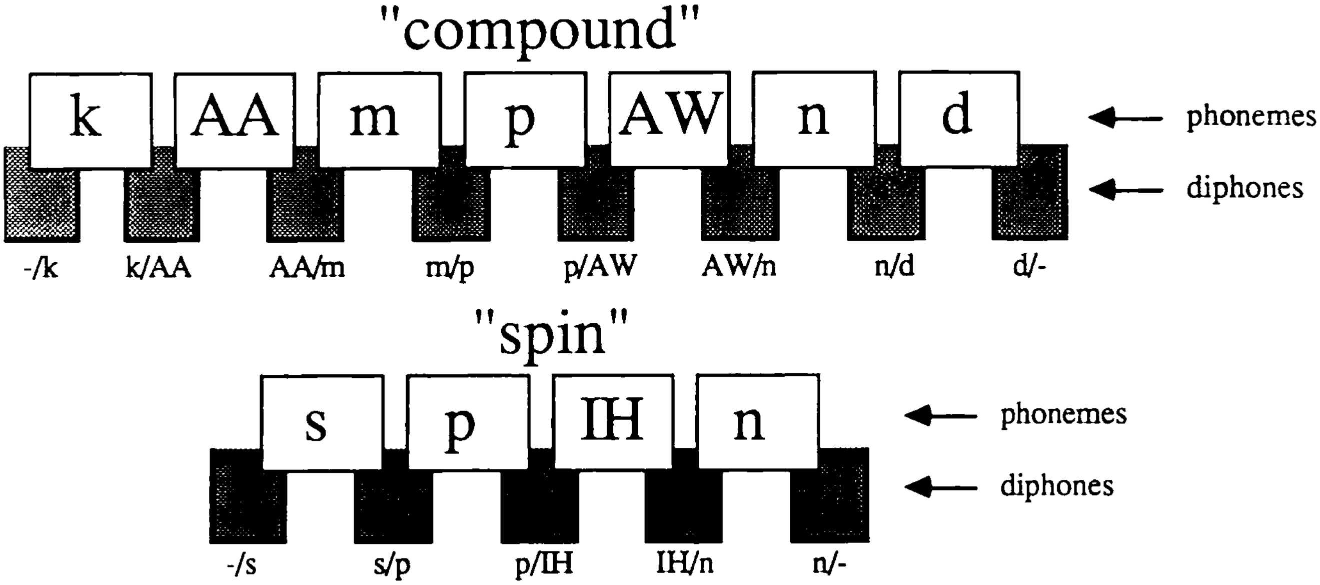
The reason that simple concatenation of phonemes does not produce acceptable speech is that the corresponding phones are not steady states, but have more-or-less smooth transitions to and from adjacent phones. Steady states, in fact, rarely occur in speech, because the speech organs move smoothly. This smoothness means that nearby phones can influence one another. This influence is called coarticulation, and it can, in some cases, have far reaching effects. Coarticulation is the

[1] Votrax is a commercial product manufactured by Votrax Inc. of Troy, New York, Votrax Inc. is a subsidiary of The Federal Screw Works and is discussed in [Kaplan and Lerner 85].

physical constraints on the vocal system imposing themselves on the speaking of phonemes.

A phone lifted out of context from a recording and then used as a representative of some phoneme will contain the transitions appropriate to the context from which it was lifted: these transitions will not work in general, but must be replaced by others appropriate to each new context in which the phone is embedded. Some synthesis systems, in fact, do something very close to this, modifying a recorded instance of a phoneme with context-dependent transitions.

However, some researchers, however [Olive 80] and [Schwartz et al. 79] for example, have chosen to concentrate on the transitions themselves. This decision is based on the observation that we can afford to be sloppy about the midpoints of the phones, if necessary, but not about the transitions, and that the best way to get good transitions is to record them. Concatenative units which are based on transitions rather than on the phonemes themselves are called diphones, or sometimes dyads. Also, demisyllables have been used by [Macchi 80] and [Browman 80] to represent similar concepts. An additional benefit from the use of diphones over phonemes is that coarticulation is well represented with diphones while no effective way has been found to do so with phonemes. In fact, many of the most successful synthesizers are based on diphone elements.



These are decompositions of "compound" and "spin" into phonemes and diphones. The phoneme decomposition is from the NRL rule system [Elovitz et al. 76]. The white boxes are phonemes and the grey boxes diphones. Note that the 'p' of both "compound" and "spin" decompose to the same phoneme. The diphones in each case are, however, very different. The diphone "m/p" has a full stop and a plosive while "s/p" is aspirated and has a very minimal plosive character.

Figure 1: Diphones and Phonemes

There is no *a priori* way to generate a list of diphones. The diphones are adjusted in time and frequency for a particular synthesis example. A particular vocabulary element therefore spans a variety of sounds. For example, "long" and "lo-o-o-ng" might be made from adjustments of the same root sound or we could have many 'o' sounds of different length. The choice of whether to add a particular diphone to the vocabulary or to adjust another element to fit is usually made in a somewhat arbitrary manner. Typically, diphone vocabularies of 800 to 2000 elements have been used ([Olive 80] and [Schwartz et al. 79]). The choice of the size of the diphone vocabulary poses a dilemma. We can use a small number of elements and use several methods to adjust the elements depending upon context or we can use a very large number of more context-dependent elements. Clearly there is some balance to be struck between having large context-dependent vocabularies and small vocabularies with lots of adjustments.

It is helpful to compare the synthesizer sections of several actual implementations of text-to-speech systems. The systems reviewed take somewhat different approaches. All of these approaches fit within a spectrum of possible solutions between very large vocabularies and small vocabularies with modifications.

1.1.1 The MITalk synthesizer

The synthesizer section of the MITalk, see [Allen et al. 79] and [Rudnicky 81], text-to-speech system concatenates phonemes. Each phoneme is stored as a single LPC frame [2] of 5 msec duration. The voice produced is a concatenation of these frames which were derived from someone's natural voice. Each phoneme is characterized by one or more of approximately 50 possible properties. These properties help to guide the synthesizer in making choices about how to adjust the phonemes.

A stream of phonemes with some punctuation and stress markers is the input to the system. The duration of each phoneme is computed, as well as the target fundamental frequency [3] where applicable. To obtain a certain length in time of a phoneme, several identical frames are abutted at the rate of 5 msec per frame. Fundamental frequency is controlled by means of adjustment of the excitation function to the stored LPC parameters. The excitation function is the input to the filter of the LPC model. In the LPC model, the parameters characterize the vocal tract and the excitation function characterizes the vocal chords (among other things). The lengths and frequency contours of each phoneme are computed using fairly complex functions of their features and adjacent phonemes.

Specific modifications of the LPC parameters are made based on rules. Also, the smoothing of the parameters is done in various ways depending upon the context. Many of the adjustments made are in the form of very specific changes to particular LPC parameters.

The image that one could construct of this method is that of having something like 42 different color bricks; each representing a different phoneme. Depending upon context we might have five green bricks in a row followed by a single yellow brick. This is equivalent to adding LPC frames for each 5 msec of the phoneme required. The bricks are then sculpted together to make a smooth and aesthetic overall fit.

The MITalk system can be characterized as being pure phonemic in the sense that the concatenative elements are steady-state phonemes. All context dependent adjustments are done after concatenation. The adjustments are controlled by means of a feature list for each phoneme which represents the physical constraints on the phoneme. The MITalk synthesizer uses very crude building blocks and very complex adjustment techniques.

[2] An LPC frame typically represents several milliseconds of speech as a list of about 12 parameters. These parameters characterize a linear filter representing the vocal tract.

[3] Fundamental frequency (F0) is basically the pitch of the speech. [Kutik et al. 83] discusses the relationship of F0 contours to English syntax.

1.1.2 The Bolt Beranek and Newman synthesizer

The synthesizer developed at Bolt Beranek and Newman (BBN) uses diphones as the concatenative elements [Schwartz et al. 79]. In addition, the method used for choosing an element is pure diphonic. The system takes a stream of phonemes and chooses which concatenative element to use based upon pairs of adjacent phonemes. Since there are 42 phonemes this gives 42 x 42 possible phoneme pairs for a total of 1764 different possible elements.

There are several problems with this approach. Not all pairs of phonemes actually occur in spoken English so the 42 x 42 matrix is actually sparse. A more serious problem is that there are a great many context dependencies in diphones that are deeper than two phonemes. In fact, there are diphone context dependencies which require as much as four adjacent phonemes to resolve. These dependencies are largely due to coarticulation. Rather than add an additional two dimensions to the table of diphones, the BBN synthesizer uses allophones. Therefore, there might be several different representations of the same diphone where the only difference is the context. This, in effect, enlarges the vocabulary from 42 phonemes to 42 plus the number of pseudo-diphones (or allophones). [Schwartz et al. 79] estimates that 2000 elements are needed for high quality speech. This figure suggests a fairly large number of special cases.

In addition to choosing concatenative elements, the BBN synthesizer does a fair amount of modification of the final speech in time and frequency. This has been a focus of work for speech processing as well as for use in text-to-speech synthesizers [Schwartz et al. 79], [Roucos and Wilgus 85b], [Roucos and Wilgus 85a], and [Roucos and Dunham 85]. The techniques used, while not specifically important to the current work, are worth mentioning because it is important to note that even with large concatenative vocabularies, processing of the final waveform is still needed.

The time-warping techniques used in the BBN synthesizer involve a model of "elastic" and "inelastic" parts of a diphone. Certain parts of the diphone are considered to be unmodifiable in time (inelastic) while other parts are considered modifiable (elastic). In addition, there is a parameter smoothing technique applied to improve transitions between adjacent diphones.

1.1.3 The Bell Laboratories synthesizer

The Bell Laboratories synthesizer is an implementation that is intended to be general. It is constructed to be applicable for a wide variety of different languages and so its structure is fairly flexible.

The concatenative element used in the Bell Laboratories synthesizer is the diphone. The context considered for the choice of an individual diphone varies from the adjacent two phonemes to the adjacent four phonemes depending upon the circumstances. While having a variable context window makes the system more complicated than a fixed context size, it allows for a broad range of possible concatenative elements.

Like the BBN synthesizer, the Bell Laboratories synthesizer uses the concept of stretchable and non-stretchable parts of voice material. Transition regions and areas of non-steady state voice are marked as non-stretchable while others are marked as stretchable using a variety of methods. Depending upon the class of the diphone, different stretch methods are applied. Some regions of diphones are stretched by repetitions of pitch periods while others are simply interpolated or other techniques are applied. The classifications of diphone regions as well as the individual stretch methods are constructed by hand.

The use of a variable context size as well as a wide variety of stretching techniques makes this synthesizer the most flexible of the three examined. The Bell synthesizer was built with the idea in mind of using the same synthesizer for other languages such as Chinese. The use of several stretch techniques as well as several techniques for element choice has resulted in very natural sounding speech. The amount of human effort required, however, in classifying the pieces of voice is great.

1.1.4 Issues to be resolved

The builder of a synthesizer section of a text-to-speech system must resolve several issues in its design. The type of concatenative element must be chosen. We have seen systems with phonemes as the concatenative elements as well as systems using diphones as concatenative elements. A method for selecting the element to be used from the phoneme stream must be devised. Some method of time-scale modification must be incorporated into the system. Finally, a system for smoothing and adjusting the final product waveform must be chosen.

Phonemes are an appealing choice for the concatenative element for several reasons. They are backed by a good linguistic model while diphones are less well defined. The number of phonemes is small and well known. Phonemes, however, must be massaged a great deal more than diphones in order to represent such acoustic phenomena as coarticulation and aspiration. Diphones, on the other hand, create a seemingly bottomless pit of different elements that are difficult to manage. Nevertheless, given our lack of knowledge of how to make concatenations of phoneme elements sound natural, a diphone based system appears to hold better hope of producing natural sounding speech. Diphones solve more problems than they create, however we need to find solutions to the management problems in a diphone system.

If we use diphones, or for that matter any concatenative element that is more numerous than phonemes, then we must come up with a scheme to choose which individual element to use in a particular instance. This problem can be characterized in terms of the width of the context dependency. For example, a particular phoneme in the context of some other specific phonemes calls for a certain concatenative element. It turns out that the context dependency ranges in width from 1.5 to about 4 phonemes. This means that depending upon the circumstance a synthesizer must look around the target phoneme in a window 1.5 to 4 phonemes wide in order to make a reasonable choice.

Once the actual concatenative elements are chosen certain large-scale considerations must be dealt with. Typically, some part of the text-to-speech system has computed a fundamental frequency contour for the sentence as well as word accents and durations. These parameters must be imposed on the stream of concatenative elements. This means, for example, that some means of stretching the element must be used along with some way to adjust the frequency and amplitude.

The real problem in this aspect of synthesis is that we know the large scale desired results but not the details. For example, we may know the desired starting frequency and ending frequency of an element but nothing of the way to get there. As for stretching, do we simply elongate all parts of an element or do we leave parts alone and stretch other parts? A real system must pick some method.

Some form of element-to-element smoothing is usually employed to prevent abrupt changes in certain parameters. Usually the smoothing is done on each of the parameters used to store the elements. Different amounts of smoothing is required for different parameters.

1.1.5 Resolution

It is useful to review some of the history of concatenative or constructive synthesis. The notion of viewing syllables as concatenated demisyllables and affixes was first introduced by Fujimura in [Fujimura 76]. Fujimura finds that English syllables may be decomposed in to "phonetically and phonotactically [4] well-motivated units" such that the total vocabulary size is about 1000 entries. Fujimura. however, recognizes that there are context related difficulties in using syllables as the concatenative base. Some of the real-world problems encountered in this approach are discussed in [Lovins 78].

From a somewhat different perspective the work of Olive in [Olive and Spickenagel 76] and [Olive 80] considers the problem of concatenative strategy from an acoustic point of view. The major point of the work is that we need to record transitions and use various forms of interpolation to generate the material in between. What is important then is to record the material that we can not recreate through interpolation. It might be suggested that the approach should be to make a vocabulary that consists only of transition region voice regardless of weather the elements are derived from phonemes or syllables.

Consider the problem of context dependency in terms of a sparse 4 dimensional table, 42 elements on each axis. The pointer to the correct concatenative element is found in the table defined by the four phoneme context. This table has 42 to the fourth or 3,111,696 entries. The sparseness is a result of the fact that only a small number of four phoneme sequences actually occur in English. The task of developing a concatenation strategy can then be considered as a task in finding the appropriate space reduction on this table.

The BBN system, for example, considers the table as being two dimensional but then selectively explodes the table with allophones. The Bell Synthesizer considers a window up to 4 phonemes wide but uses algorithmic means to narrow the window in most cases.

Even with this model of context dependency, smoothing is still required. The point of all of this is that limitations in the physical systems give rise to complications in the production methods that may have nothing to do with semantics. The whole methodology of concatenative synthesis in text-to-speech systems is intended to bridge the gap between the somehow semantic realm of phonemes to the physically constrained domain of actual speech.

The problem of stretching the elements in time seems to be getting close to solution. Time scaling is best done with a system which recognizes that different parts of an element need to be dealt with in different ways. Voiced sections can be stretched by adding pitch periods while plosive sections are best left untouched. This approach adds another burden to the concatenative vocabulary builder: each element must be marked for the different types of stretching permitted.

The capabilities of the human ear are far greater than is required for speech communications. The ear then becomes a very sensitive detector of defects in production method that have nothing to do with content. To build synthesizers that will please the ear the physical limitations of the vocal system must be considered in the production methods.

[4] Phonotactics refers to the mechanical motions of the vocal organs during the process of producing speech.

1.1.6 Articulatory synthesizers

There is a class of speech synthesizers known as articulatory synthesizers which by their construction attempt to mimic the vocal tract and muscles [Coker 76]. While this would seem to hold great promise of producing natural speech, there are some severe problems with this approach. Articulatory synthesizers account for the various spectral complexities of natural speech using a model of articulatory dynamics and control. [Coker 76] presents a fairly complete model of the human speech process. This model examines the constraints on the vocal system and produces output which has similar spectral properties to natural speech.

The input to the synthesizer is in the form of low-level muscle actions and tract configurations. This puts us in the same position as before: concatenation of groups of muscle actions instead of a concatenation of sounds. We also have very little to go on in the creation of these commands to the synthesizer as the transformation from muscle actions to speech sounds is unknown. The model constrains the output but provides little guidance in generating input to the system.

Matters are made even worse by the observation that the vocal tract may not be the ideal organ for making sound. In fact, the vocal system is fairly close to its limits when we speak [Liberman 87]. This means that the articulatory synthesizer is ill-conditioned in the sense that small input errors result in gross output errors. The fact that the physical properties of the vocal system have such a profound effect on the character of the voice produced should provide a great deal of help in generating natural sounding speech. This dependence is so detailed, however, that our knowledge of the biophysics involved is inadequate for useful high-quality speech production with these types of synthesizers.

1.2 Choice of elements

Once the concatenative strategy or strategies have been chosen, elements to be concatenated must be acquired. In all cases of interest, the elements to be used in the synthesizer are extracted from natural human speech. Since one natural realization of a diphone must stand for all the possible realizations of that diphone, the choice is rather important. To date, the extraction of these elements has been by hand with only *ad hoc* optimization (with respect to the quality of the speech) attempts. The manner of choosing the actual piece of voice that will be used to generate speech is not well understood and is the subject of much of the proposed research in this thesis.

It is one thing to decide what kinds of elements are going to be used (diphones or phonemes) but it is still another to pick them out of a stream of natural speech. Since, in the case of diphones, we are picking out transition regions we must be careful to capture the whole transition. One can think of speech as being a series of transitions and steady states. The choice of how much of an element to capture is very important. In the case of plosives, for example, the attack and decay regions are very significant.

Elements must be captured in the context in which they are intended to be used. Some method of detecting that context must be used. To date this has been done by simply listening to the speech to find the appropriate regions.

Generally, the extracted elements are taken from unstressed positions in a word. This is done because stressing causes changes in the element which are fairly independent of the element type. The stressing is dependent on position within the word and word morphology. Therefore we can stress a part of a word from unstressed material more easily than we can make an unstressed part of a word from a stressed element.

In this thesis we approach the problem of concatenative vocabulary selection as the optimization of a codebook with respect to some quality measure. The measures used will be a combination of established distance measures. The following section discusses the methodology.

2 Method

The experiments proposed have several parts in common. There will be a fixed body of natural voice material as input to the experimental system. There will be a part of the system which selects pieces of the input set for use by the synthesizer. The synthesizer will then rebuild some subset of the input material and then another part of the system will evaluate, using as a distance measure the generated speech against the natural input speech.

2.1 Limited domain

A fixed body of speech signals will be used as the target to be modeled in the optimization. This fixed body will be made large enough to have a significant number of each proposed concatenative element. The sample material will be natural voice recorded under high quality conditions and stored in digital form appropriate to the synthesis methods used. In fact the use of high quality conditions is not as important as consistent conditions; this is so concatenated pieces have the same background noise. The sample will be annotated to indicate candidate regions using a waveform editor.

The sample domain to be modeled will exclusively contain consonant-vowel-consonant (CVC) words. This may be expanded later; however, there are significant advantages to using CVC words. It is possible to construct a great many sentences using only CVC words so the input speech will be meaningful text. It should be pointed out that the problem is decomposable to small vocabularies. If the experiment is run using a subset of phonemes then an optimized concatenative vocabulary will be generated for that subset. Later, if more phonemes are added then the other phonemes will still function as before; of course, they may no longer be optimal over the new, larger set.

CVC words are stressed in only one way. This means that stressing does not have to be extracted from candidate concatenative elements. While it is possible to pick neutral stress elements and then use other techniques to add stress, this is an added complication that is not significant in the proposed work [Schwartz et al. 79].

Since the synthesis of CVC words involves the concatenation of a fairly small number of elements, the size of the test set is kept small with the use of CVC words. The problem is fairly decomposable so there is no loss of generally in this approach. Further, we conjecture that the elements chosen with the smaller word size will not significantly differ with the use of longer words. Further, it is found that in a large word list of 250,000 English words that 9300 are CVC and in a smaller list of 35,000 common words that 3300 are CVC. CVC words are a significant fraction of English words.

One approach taken will be to use a single voice. A particular choice will be made by hand of concatenative elements for the synthesizer and then the choice will be optimized within that voice. The optimal choice is defined to mean the choice of concatenative elements which minimizes the cumulative distance between the synthesized words and their respective natural words in the sample speech.

It is not necessary for the purpose of evaluating the approach to construct complete sentences using CVC words. One word in a consistent place in the sentence will suffice. As a first example, a sentence was constructed that has the desirable property that it isolates the CVC word from coarticulation effects. To achieve this isolation, the phrase prior to the CVC word ends in a full stop such as /t/ and the phrase after the CVC word begins with a plosive such as /p/.

The sentence used in the first trial was "The fast [CVC] passed the store." This sentence does not always make sense but is always grammatical, where the [CVC] is a noun. For the first trial, three initial, middle and final phones were chosen. Since it is desirable to use only real words it is important to choose the nine elements carefully.

The Websters word list was passed through a NRL rule [Elovitz et al. 76] based text-to-phoneme program. The result was then filtered for CVC words and then the resulting set was evaluated to find the list of three initial, middle, and final phonemes which produces the largest number of real words. The result was /b/, /f/, and /m/ for initial phoneme; /EH/, /AE/, and /IY/ for middle phoneme; and /d/, /n/, and /t/ for final phoneme. This resulted in the following list of words:

word	phoneme	part-of-speech
bed	/b/EH/d/	noun
ben	/b/EH/n/	noun
bet	/b/EH/t/	verb, noun
fed	/f/EH/d/	verb, noun (slang)
men	/m/EH/n/	noun
met	/m/EH/t/	verb, noun (Baseball)
bad	/b/AE/d/	adjective
ban	/b/AE/n/	noun, verb
bat	/b/AE/t/	noun, verb
fad	/f/AE/d/	noun
fan	/f/AE/n/	noun, verb
fat	/f/AE/t/	noun, adjective
mad	/m/AE/t/	adjective
man	/m/AE/n/	noun
mat	/n/AE/t/	noun
bean	/b/IY/n/	noun
beat	/b/IY/t/	noun, verb
feed	/f/IY/d/	noun, verb
feet	/f/IY/t/	noun
mead	/m/IY/d/	noun
mean	/m/IY/n/	noun
meat	/m/IY/t/	noun, verb

A list of CVC words with their part-of-speech. The words are derived from the phonemes /b/, /f/, /m/, /EH/, /AE/, IY, /d/, /n/, and /t/.

Figure 2: A list of CVC words

The script consisting of the sentence "The fast [CVC] passed the store" with each of the above words used in turn was recorded using the author's voice.

2.2 Recording Techniques

Scripts are digitally recorded on a personal computer. The speech is low-pass-filtered to exclude frequencies above 10khz. The filter serves as an anti-aliasing filter for the 20khz, 12 bit digitizer on the personal computer. Since it is difficult to create an analog filter that has both a sharp cut off frequency and a flat response in the region just before the cut off, the signal is digitally refiltered to 5khz. Thus, the analog filter serves to prevent aliasing while the much sharper digital filter provides a flat response and acts as an anti-aliasing filter so that the digital signal may be digitally down-sampled to a 10khz sampling rate. Approximate markings of the diphones of interest are made using a simple voice editing facility on the personal computer.

2.3 Distance measures

Speech distance measures are used to evaluate the differences between two pieces of speech. The difference between two pieces of speech only has meaning with respect to something. Just what that something is depends upon the application. Speech distance measures are used in speech recognition, speech synthesis, speech coding, and speaker verification. Each of these domains demand different things from the distance measures that they use. Generally, measures involve a model of distortion; that is the speech examined is assumed to have been distorted by some process. Two segments of speech are compared to see how much distortion would have to be present to make one look the same as the other. Obviously, the distortion model as well as the comparison techniques used depend upon what one expects to find.

There are several properties that most useful distance measures have.

1. $d(x, y) = d(y, x)$ symmetry
2. $\left.\begin{array}{l} d(x,y) > 0 \text{ for } x \neq y \\ d(x,x) = 0 \end{array}\right\}$ positive definite

There is the triangle inequality which may sometimes be useful and which makes $d(x, y)$ a metric.

3. $d(x, y) \leq d(x, z) + d(y, z)$

In addition, the following properties are usually desirable in a distance measure.

4. $d(x, y)$ should have a physically meaningful interpretation
5. $d(x, y)$ should be efficiently computable

A distance measure may be thought of as a transformation that maps the speech segment into a parameter space. A distance measure is, however, not a general transform. Many different speech samples may map to the same place in the parameter space. The parameter space may be many or infinite dimensional.

A geometrical distance measure is then directly applied to the parameter space. In the absence of a reason to use another method, Euclidian distance is frequently used: $d(X, Y)^2 = \sum_i w_i(x_i - y_i)^2 \mid w_i = 1$. Euclidian distance is, however, almost always the wrong geometrical distance measure to use in the parameter or feature space. The reason for this is that in order to use Euclidian distance, the dimensions of the feature space must be orthogonal. In general this is not the case.

In fact the topology of the feature space can be extremely complex. However, statistical analysis of the features can help in determining weights (w_i) used for the sum. Of course a linear sum itself may not be the right thing to use as the weights may vary depending upon location.

Another type of measure is a deformation measure such as dynamic time warping distance. Dynamic time warping (DTW) is a process by which the time axes of two waveforms are selectively deformed so as to find the best possible alignment between the signals. It is possible to find the minimum deformation required to bring the two waveforms into maximum alignment according to some distance measure. The total deformation required can be regarded as a measure of the similarity between two waveforms with respect to the distance measure used. In addition, DTW is frequently applied before the application of other measures. DTW is in this way used to normalize for rate-of-speaking differences. The dynamic programming techniques used in DTW can also be applied to the frequency domain with certain constraints. Dynamic frequency warping is frequently used as an attempt at speaker normalization.

2.3.1 Specific distance measures

There are several different measures that we will consider here. In each case we shall present the measure and an interpretation of that measure. The measures considered are: log spectrum, cepstrum, log likelihood, Itakura-Saito, dynamic time warping, and dynamic frequency warping.

All of the measures which involve transformations to frequency domain make some rather strong assumptions about the stationarity of the signal. Voice, in general, is not stationary. The measures are applied over time intervals which are short enough so that the signals are approximately stationary over their duration.

Spectral measures Consider two spectral models $\sigma/A(z)$ and $\sigma'/A'(z)$. These models refer to σ, the half sampling rate of the signal and to $A(z)$ which is the transfer function of the filter that represents the signal over the time region considered. The $A(z)$ function is formally defined later in this section. The stationarity assumptions are important here because $A(z)$ is really only defined for stationary signals. The difference between these models on a log magnitude versus frequency scale is

$$V(\theta) = \ln[\frac{\theta^2}{\mid A(e^{i\theta}) \mid^2}] - \ln[\frac{\theta^2}{\mid A'(e^{i\theta}) \mid^2}],$$

where θ is normalized frequency. $V(\theta)$ is a very important function for comparing spectral models since in the log-spectrum domain spectra add. This creates a space in which to consider distances. One such distance may be defined as:

$$(d_p)^p = \int_{-\pi}^{\pi} \mid V(\theta) \mid^p \frac{d\theta}{2\pi}.$$

When $p = 1$, the measure returns the mean absolute log spectral measure and when $p = 2$, the measure returns the mean-square log spectrum distance. For $p \rightarrow \infty$, the peak log spectral difference is obtained.

The choice of what value of p to use depends in part on what kinds of differences are anticipated. From the example presented in [Gray and Markel 76] it is clear that frequency alignment before the application of the measure may be beneficial. [Gray and Markel 76] goes into great detail on the subject of the meaning of the choice of p. Essentially, the higher the value of p the less small differences contribute to the measure. An interesting result concerning the choice of p is reported in [Barnwell 80]. In this work, objective quality measures are compared to the subjective Diagnostic Acceptability Measure (DAM) [5]. It is found that values of $p = 4$ to $p = 8$ yield a higher degree of correlation to the DAM than any other value for p.

One physical interpretation of this measure is perceptual. The inner ear is known to be essentially a frequency-analysis device; hence the use of a spectral model. In addition, the nerves in the ear are known to have a partially logarithmic response; hence the log of the spectrum.

Cepstrum measure The cepstrum is defined as the Fourier transform of the log of the power spectrum [Parsons 86]. If the cepstrum is short-pass filtered (or "liftered" as it is called in the cepstrum domain),then the result is a smoothed spectrum. This is referred to as cepstral smoothing. The cepstral measure is defined as:

$$[u(L)]^2 = \sum_{k=-L}^{L} (c_k - c'_k)^2 = (c_0 - c'_0)^2 + 2\sum_{k=1}^{L}(c_k - c'_k)^2,$$

where the c_i are the cepstral coefficients defined by:

$$\ln[A(z)] = -\sum_{k=1}^{\infty} c_k z^{-k},$$

with c_0 and c'_0 just gain terms.

With this measure, $u(L)$ can be interpreted as the RMS (root mean square) distance between the log spectra after each log spectrum has been cepstrally smoothed to L coefficients. The measure is positive definite as long as L equals or exceeds the filter order of the input speech [Gray and Markel 76]. The cepstrum measure is really a variation on the spectral distance since the underlying transformation is still into the frequency domain. In fact, as L goes to infinity $u(L) = d_2$, where d_2 is the spectral distance measure defined previously [Gray and Markel 76].

Itakura measure Speech is frequently coded with a predictive system such as linear predictive coding (LPC). The predictive system always has an error sequence the power of which is minimized for the predictor. For LPC the prediction error series is just:

$$e(n) = \sum_{i=0}^{M} a_i x(n-i).$$

So with $a_0 = 1$, the total squared error is

$$\alpha = \sum_{n=-\infty}^{\infty} [e(n)]^2.$$

[5] The Diagnostic Acceptability Measure is defined in [Voiers 77].

One way of looking at this type of coding of speech is to view it as a smoothing filter. The transfer function of that filter is given by:

$$A(z) = 1 + \sum_{i=1}^{M} a_i z^{-i}.$$

This filter minimizes the error α for the series $x(n)$. If the series $x(n)$ is passed through an inverse filter

$$A'(z) = \sum_{i=0}^{M} a_i' z^{-i}$$

which minimizes the error α' for some other data series $x'(n)$, then the total squared energy of $x(n)$ through $A'(z)$ is called δ and is given by:

$$\delta = \sum_{n=-\infty}^{\infty} \left[\sum_{i=0}^{M} a_i' x(n-i) \right]^2 .$$

This gives us an interesting way of comparing two predictors in terms of their prediction errors. The ratios δ/α and δ'/α' are referred to as likelihood ratios. A clearer way to look at the likelihood ratios indicates that they are just a way of summarizing the differences between two predictors over all frequencies as can be seen from this integral (from [Gray and Markel 76]):

$$\delta/\alpha = 1 + \int_{-\pi}^{\pi} \frac{\mid A'(e^{i\theta}) - A(e^{i\theta}) \mid^2}{\mid A(e^{i\theta}) \mid^2} \frac{d\theta}{2\pi}.$$

[Itakura and Saito 70] develop the integral

$$\Xi = \int_{-\pi}^{\pi} [e^{V(\theta)} - V(\theta) - 1] \frac{d\theta}{2\pi},$$

where $V(\theta)$ is the difference in spectral models developed above. This measure is more easily computed with the likelihood ratios thus:

$$\Xi = (\sigma/\sigma')^2(\delta/\alpha) - 2\ln(\sigma/\sigma') - 1.$$

In [Itakura and Saito 70] the authors state that they believe that their measure considers perception (spectrum) as well as generation (predictor) issues. In addition, [Itakura 75] shows that the Itakura measure is optimal from certain points of view for comparing linear predictive models. In fact the Itakura-Saito measure has stood the test of time as a dominant distance measure in speech recognition work.

Dynamic time warping distance Dynamic time warping is a dynamic programming algorithm for the computation of the minimum deformation of one time series to match another. One can look at DTW by thinking of placing two time series on perpendicular axes. A path is drawn in the plane created which is a curve that creates a correspondence between the points in the two series.

For a given distance measure the DTW algorithm finds the deformation curve which minimizes the total deformation needed to make one series into the other. The total distance is referred to as the dynamic time warping distance and is given by:

$$D(R,T) = \min_{c} \left[\frac{\sum_{k=1}^{K} d(c(k)) \cdot w(k)}{\sum_{k=1}^{K} w(k)} \right],$$

where $c(k)$ is a parametric curve in the plane, $w(k)$ are weighting coefficients, and $d(c(k))$ is the distance measure on the curve. R and T are the reference and test series respectively. This explanation of DTW can be found in a great many references including [Charot et al. 86].

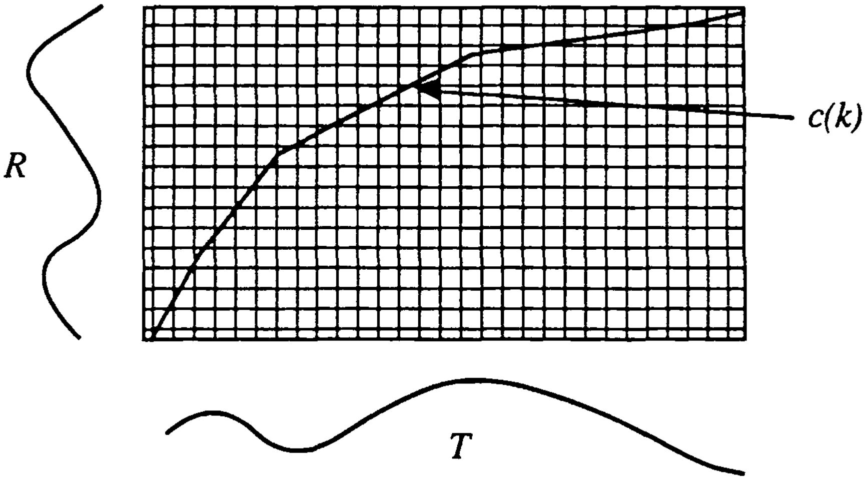


Figure 3: An illustration of dynamic time warping

Dynamic time warping is a technique which has been found to be widely useful in speech processing. Most of the other distance measures address low-level issues in either perception or generation. On the other hand, dynamically warping the time scale seems to hold promise of normalizing for differences in higher-level processes. For example, if the only difference between two words is their rate of speaking, then we would like a process that would normalize them to be nearly the same. Dynamic time warping comes close to being able to perform this normalization. Other measures applied after DTW would then seem to have more significance.

Dynamic frequency warping (DFW) distance A distance measure based on non-linear frequency warping is discussed by [Matsumoto and Wakita 79] and applied to speaker normalization in speaker independent speech recognition. The measure uses the minimum mean squared distance over all possible choices of frequency warping functions. The minimum warping function is obtained in much the same way as for dynamic time warping. However, a variety of constraints on the warping function must be applied. The actual distance measure is:

$$d(S',S) = \min_{w} \left[\sum_{k=1}^{N} (S'_{i(k)} - S_{w(i(k))})^2 \frac{W(w(i(k)))}{\sum_{k=1}^{M} W(w(i(k)))} \right],$$

where $w(i(k))$ is a frequency warping function and $W(w(i(k)))$ is a weighting coefficient.

Since it is assumed that the frequency shifts are mostly due to differences in vocal tract length, constraints can be set on the warping. In the $f' - f$ plane, the warping function is restricted to the region bounded by two straight lines whose slopes are l_S/l and l_L/l, where l_L is the maximum vocal tract length, l_S is the minimum vocal tract length, and l is the actual vocal tract length of the sample speech. Any warping function that crosses these lines would correspond to a non-physical distortion of the spectral characteristics.

Glottal [6] differences are compensated for by normalizing for spectral slope. A least squares line fit is applied to the spectrum on the assumption that the log of the observed spectrum is the sum of the vocal tract transfer function and the glottal source spectrum with a relatively minor contribution from radiation efficiency of the lips. The resulting spectrum is then just the vocal tract transfer function.

The result of these two techniques (spectral slope normalization and DFW) is an improvement in speaker independent word recognition rates. For a male voice, the recognition rate went from 78.7% to 85.2% with both techniques applied. For a female voice, the rate went from 64.8% to 84.3%. This normalization technique (or measure) clearly is adjusting for a great many of the speaker differences present in speech.

2.3.2 Correlation to subjective measures

[Gray and Markel 76] relates many of the distance measures to what is referred to as just noticeable differences (JND). For example, in a certain measure the just noticeable difference for vowels is 1.5 dB and for unvoiced speech 0.4 dB. The just noticeable part is subjective and the measurement is objective. This technique allows some calibration of a distance measure to a human perception threshold.

An early survey of speech quality measures is [Hecker and Guttman 67]. The major concern in this work is listener preferences in the presence of a variety of distortions. Isopreference contours are drawn in a paired-comparison paradigm. The scale is expressed in "Transmission Preference Units" (TPU). Isopreference contours are shown with loudness and noise level on the X and Y axes and TPU as the Z component.

A much more sophisticated approach to relating subjective and objective quality is taken in the work of Barnwell [Barnwell 79]. In this work, the author develops criteria for the evaluation of objective quality measures. Objective quality measures are simply defined as methods of measuring speech signals that do not require human subjective assessment.

[Barnwell 79] observed that "[distance] measures which do not use semantic, syntactic, and other language related information cannot correctly predict the quality of a speech coding system", it is noted that the class of distortions introduced during speech coding are not correlated to semantics. [Barnwell 79] develops methods of evaluating the quality of distance measures used for judging coding systems. These measures are called fidelity measures. The results of comparing a variety of objective measures to a paired acceptability test [7] [Voiers 77] are presented.

[6] Glottal refers to the slit-like orifice between the vocal cords [Flanagan 72]. The term is often used to refer to whole of the excitation mechanism in voiced speech sounds.

[7] The paired acceptability test is a subjective speech quality test involving a human jury.

2.3.3 Significance of distance measures to synthesis

It is reasonable to examine the significance of these distance measures with respect to the problem of evaluating high-quality speech. It is obvious that all of these measures converge to zero distance as the speech signals get very close. That is to say that as any one measure approaches zero, so do they all. Further, when two speech signals are completely different, the measures generally return some number that is probably meaningless. Dynamic frequency warping may simply fail in this case due to exceeding the constraints on the warping function. It is helpful to look at the domain in which these measures have been the most useful – speech recognition.

These distance measures can be thought of as creating a mapping of a segment of speech to a multidimensional region. If two segments of speech are both within the same region then they are really indistinguishable from the point of view of the measure. If we try to have a great many regions, then we find that they will start to overlap. This is one way to understand the fact the speech recognition with almost any of these measures can be performed for about 100-200 words to reasonable accuracy. Going beyond about 200 words (with only measures and no semantic or statistical information) is very hard. This is because the 200 regions overlap so severely that it is impossible to tell reliably which region the speech segment falls into.

The speech measurement process can be thought of as a compression of a speech segment to a single number. Obviously, the point is to throw away details in the speech which are irrelevant to what the measure is measuring. Different measures throw away different things. Most of the measures, however, group segments (in different ways) according to their spectra. This is fine for differentiating between say a violin and a piano. In that case the spectra are very different.

Unfortunately, the evaluation of speech synthesis is a different problem. Consider that we already know that we are listening to a violin and maybe even what notes are to be played. The question is now: how good is the playing? or what is the quality of the instrument itself? Clearly this is a hard problem; but, more importantly, it is a different problem from the recognition problem for which the measures have been used before. Further, the measures that have been applied to speech quality evaluation have only had high correlation to human opinion when the distances are very small and are really only good for determining intelligibility rather than quality.

The situation is not completely hopeless, however. At the very least we can always fall back on the best measure of speech quality – a human jury. Human juries however often return information that is difficult to interpret from the perspective of improving the synthesizer. In addition, the distance measures examined seem to have the wrong characteristics for synthesis evaluation when applied to speech in general. It seems reasonable to believe that applying these measures on very small pieces of speech to look at individual events may give better sensitivity to speech quality. This, however, remains to be seen.

2.4 Inner optimization loop

The inner loop of the optimization procedure consists of choosing a model for the sample speech and then evaluating the quality of the model.

Choosing a model involves the selection of an element for each piece of the vocabulary from the input sample. The input sample will contain several possible candidates for each concatenative element in the vocabulary. The concatenative elements are rescaled in time for different uses. Therefore, an additional aspect of the choice of a concatenative element is length, in time, of the vocabulary element.

The search pattern through the space of possible models will, ultimately be important. However, the early experiments will involve small enough samples that the entire space of possibilities will be spanned. In later, more extensive, work a heuristic search technique may be necessary. The search pattern used will be dependent on the kind of information available from the measures used. Since it is not clear yet what the measures will yield, further speculation on this aspect of the work is not appropriate. In addition, the option is still available to do exhaustive searches.

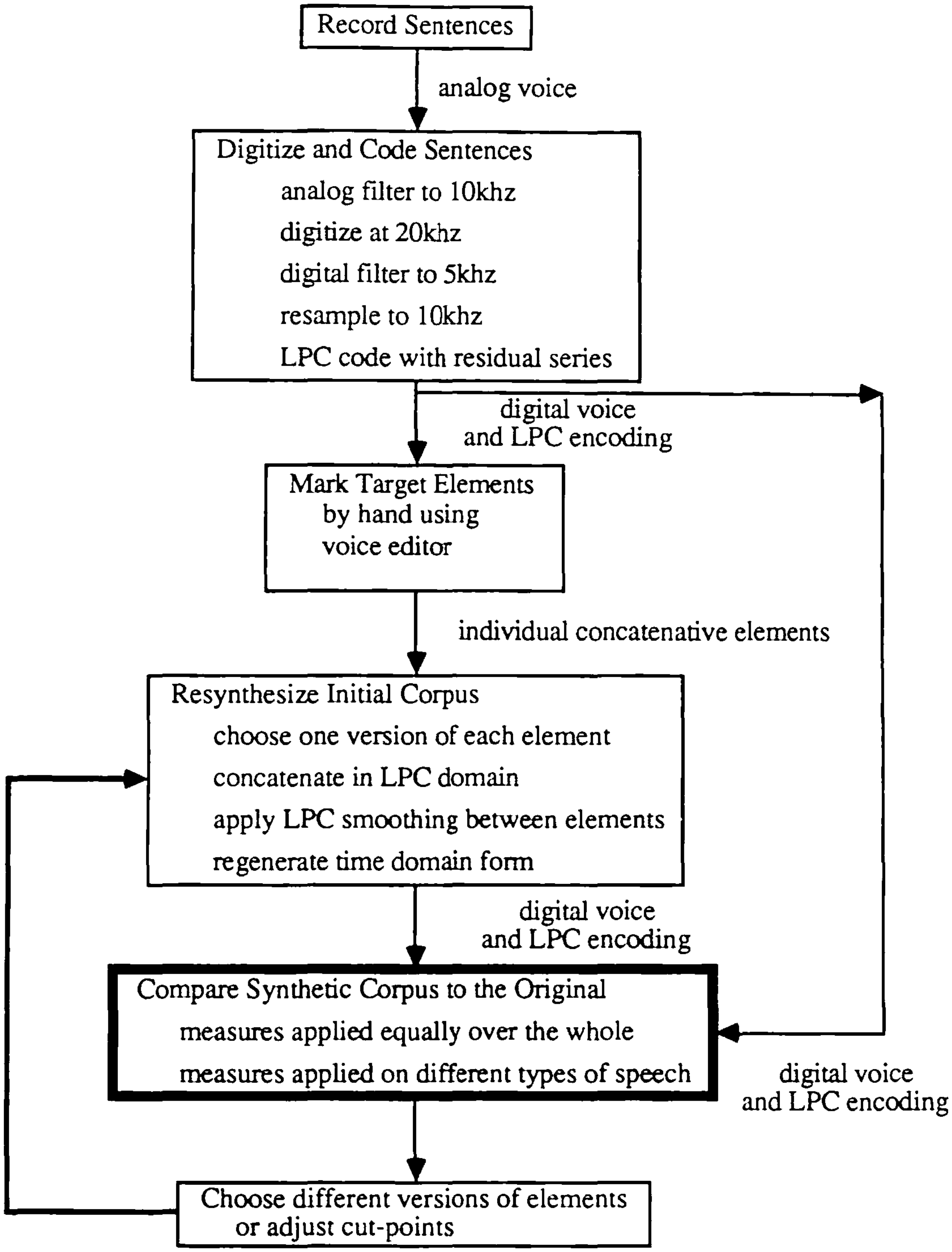
The initial experiments with the nine phoneme inventory consist of trying to make certain inventory selections. In this small version of the inventory, we can explore some of the major issues of inventory selection. For example, there are six words in our script which contain the phoneme /EH/. Which one do we choose if in fact we permit ourselves only a single version of /EH/?

If we choose a diphonic concatenation scheme, then we will be looking for the diphone /b/-/EH/ for example. In this case we have three examples in the word list (bed, ben, and bet). Each word in the list is also recorded several times. The task that we have is to decide which actual piece of voice to use as the concatenative inventory element. In addition, for each time a word is recorded we may choose the cut points or boundaries of the element differently. Each variation on the extracted element is considered a different version.

We generate all of the candidate elements for each of the positions in the inventory. This is done with a simple digital voice editing tool. All of the possible reconstructions are then automatically assembled using each possible element choice. This step generates a great deal of reconstructed voice. However, the reconstruction is done automatically and is deterministic so we can recreate the voice and throw it away as needed to trade time for space.

The next step is to evaluate all of the possible reconstructions to find what might be considered the best. We will use combinations of different distance measures for the evaluations. In particular since we build the output ourselves, we know what sorts of speech signals to expect at different points in the speech to be evaluated. This knowledge will permit greater fine tuning of the distance measures than has been generally possible before. Extensive experimentation on this step has not yet been done. This will constitute the major work of the thesis.

Different measures may be applied to different types of words. The scope of the comparison will likely be words; however, larger or even smaller scopes may be employed. Since the comparisons will be between utterances which are assumed to be very similar, dynamic time warping (DTW) will be the starting point for constructing measures. DTW distances of derived time series such as format frequencies will also be tried. The speech recognition literature will be consulted for other promising measure choices. Of course, general listening tests will be performed to assure that the measures are in some way meaningful.


Record Sentences
analog voice
Digitize and Code Sentences
analog filter to 10khz
digitize at 20khz
digital filter to 5khz
resample to 10khz
LPC code with residual series
digital voice
and LPC encoding
Mark Target Elements
by hand using
voice editor
individual concatenative elements
Resynthesize Initial Corpus
choose one version of each element
concatenate in LPC domain
apply LPC smoothing between elements
regenerate time domain form
digital voice
and LPC encoding
Compare Synthetic Corpus to the Original
measures applied equally over the whole
measures applied on different types of speech
digital voice
and LPC encoding
Choose different versions of elements
or adjust cut-points

This is an overview of the system proposed. The boxes contain the action of one or several programs while the arrows are labeled with the kind of data that flows between the programs. The original work of this thesis is primarily in the "Compare Synthetic Corpus to the Original" box.

Figure 4: System Overview and Dataflow

3 Importance

The major impact of this work will be the creation of techniques to automatically generate concatenative vocabularies. Unless automatic or at least in some way objective techniques are used to evaluate voice it is impossible to reasonably compare the low-level quality of speech systems. As long as human judgment and patience are the sole factors in vocabulary selection, it will not be possible to compare the results of different systems independent of the voice editing.

There are several reasons to expect that the techniques discuss will produce better results than voice editing alone. Voice editing techniques permit the human editor to listen to elements in a number of contexts limited by human patience. In fact several tens of contexts are a practical limit. An automated system could examine far more contexts than a human. There may also be complex trade-offs to which the human editor may not be sensitive. Changing one element may influence the way that another element sounds in a different context as some contexts may be up to four elements wide.

3.1 The quality of synthesizer output

There are many factors which influence the subjective quality of text-to-speech systems. Many of these factors are reviewed in [Polish 86]. Certainly the morphological decomposition of the words is important. However, the words in all systems are ultimately created by the concatenation of voice elements in a synthesizer. The morphological decomposition as well as the results of higher-level analysis, such as fundamental frequency contour, serve as input to the concatenative synthesizer. All of these elements contribute to quality.

The pronunciation problem and the speaking problem are separate. A great many synthesizers pronounce words correctly only to speak them poorly. It is possible, in synthesis-by-rule systems, to identify mispronunciations or errors in decomposition and correct the rule base. Very little research, however, has been directed at optimizing the concatenative vocabularies. The need for improvement in the synthesizers is constant regardless of the higher-level decomposition system employed. A variety of different synthesizers have been experimented with, some of which are reviewed in a prior section, however, optimization of the vocabularies themselves has rarely been discussed.

All unrestricted text-to-speech systems have the property that they will make an attempt at speaking any input. The resultant output will sound different from that of other systems given the same input for a variety of reasons which will vary from system to system. In all cases, the quality of output will depend, at least in part, upon the concatenative vocabulary used and how well it is built. The concatenative strategy and vocabulary make an important contribution to the system quality because together, the strategy and vocabulary, create the set of possible sounds from the system.

3.2 A method for the creation of different voices

Synthesizers ultimately sound like the person from whom the concatenative vocabulary is taken. Since these vocabularies are built by hand and optimized in a very subjective manner, it is very time consuming to create new vocabularies in new voices. In addition, since the optimization is by hand, it is very difficult to compare the results obtained from different speakers. The creation of a means to easily build new concatenative vocabularies will facilitate research on the effectiveness of different voices in synthesizers.

In principle one would like to be able to create a text-to-speech system which could be adjusted or modified to speak in any voice desired. It would be interesting to be able to try a range of different voices using the same underlying system. By creating an automatic or at least semi-automatic means of creating, evaluating, and optimizing different sets of elements in different voices it will be possible to make new vocabularies with comparative ease.

3.3 A systematic method for analyzing the quality of vocabularies and strategies

While it is fairly straightforward to create a vocabulary which renders intelligible speech, it is very difficult, with existing techniques, to know the best way to improve the vocabulary. In any speech synthesis system, it is necessary that a single vocabulary element stand for a wide variety of sounds. The stretching and scaling processes applied improve the versatility of a particular element. However the altering of the elements also complicates the problem of determining how well the element contributes to the overall sound since, once altered, the element is really a set or class of possible sounds.

It is important to be able to make a reasonable judgment as to when it makes sense to add a new vocabulary element. In particular, decisions about when elements should be changed need to be repeatable and based on objective analysis. Given a fixed input set, it will be possible to assess the improvement from changes in the synthesizer.

3.4 Development of local acoustic distance measures

A globally applicable voice quality measure that works for all natural voices in all likelihood does not exist. However, measures which work well with pieces of speech which can be assumed to be acoustically very close would still be very valuable. The optimization which will be performed in the proposed work will be based on local distance measures.

It will be possible to examine very large samples of (the same) voice to determine variance within the measure. This will lead to better understanding of the kinds of measures which could be used for speaker dependent speech recognition. In addition, an environment for examining and testing of various speech measures would result from the work.

3.5 Better understanding of the fundamental conserved quantities in concatenative elements

Since prior work has concentrated on one voice at a time, researchers have had to take on faith that the particular concatenative strategy chosen has some universality across some set of speakers of a particular language. It is important to be able to compare the concatenative vocabularies generated for different speakers. Ultimately it would be very useful to identify those acoustic quantities which are conserved from speaker to speaker of the same element. This work will also be helpful in speaker independent speech recognition because acoustic quantities which are preserved from speaker to speaker are exactly the kinds of information needed in speaker independent speech recognition.

3.6 Concatenative synthesis as a codebook method

Concatenative synthesis may be thought of as a form of codebook [8] coding. In concatenative synthesis we are looking for a set of elements of speech to use as a codebook with which to reconstruct the original speech with minimal error compared to the original. Speech synthesis is different from conventional codebook techniques in that the concatenation technique itself involves smoothing and stretching and is generally flexible.

The advantage of viewing concatenative speech synthesis as a codebook technique is that this kind of technique has been extensively studied for transmission of ordinary digital data. If the techniques discussed in this proposal prove workable, then we will have a method of evaluating how good a codebook a particular inventory and concatenation scheme is.

[8] See [Parsons 86] for an introduction to the use of codebooks in speech processing.

4 Judging Success

Judging the degree of success of the proposed work will involve examining the range of systems that will be developed as well as results that will be demonstrated. Below are elaborations on just what we plan to develop and what we plan to demonstrate. In addition, there is a section on what sorts of criteria should be used to evaluate the final results of the work.

4.1 What we will develop

The proposed thesis work involves the development of several systems. We will be developing a set of speech distance measures to be used to guide the optimization of the concatenative vocabularies. In addition, we will be developing a set of speech manipulation tools to aid in the extraction and processing of speech material. Both of these systems have been discussed in prior sections and are elaborated below.

4.1.1 Mixed distance measures

In this research we will develop one or several measures to compare synthesized speech to other synthesized speech. These measures will be unusual in that different measures will be applied to different types of speech signals. For example, unvoiced sections of speech will be compared differently from voiced sections of speech signal.

In effect we will be building a suite of speech measures. The suite will be made up of a variety of commonly used speech measures such as those discussed in a prior section. The measures that we develop should be evaluated within the context of the concatenative vocabulary extraction system that we will build.

4.1.2 Speech extraction and manipulation tools

The voice measures that we develop will be used to evaluate candidate concatenative elements for speech synthesis. Since we will be using an optimization process to pick the best elements, we will be developing a limited synthesizer as well as vocabulary extraction tools. The ultimate test of these systems will be whether they produce speech of comparable quality to other, manually produced systems.

4.2 What we will demonstrate

In order to indicate the success and usefulness of the system and measures developed, we will demonstrate several tangible things. In particular we will demonstrate the semiautomatic extraction of a concatenative vocabulary for use in a synthesizer. We will also compare, using a human jury, the results of this work with other synthesizers and with a set of elements chosen by hand.

4.2.1 Limited domain synthesis

We will demonstrate the synthesis of a limited subset of English using a subset of English sounds. Certain sentences will be selected with individual words chosen as the target of the synthesis. The remainder of the sentence will be subjected to the same coding techniques as the concatenative elements so that the basic signal quality will be similar to the concatenated words.

4.2.2 Other voices

We will synthesize the same sentences in at least two different voices using the same underlying techniques. Concatenative elements will be generated for at least two different voices using the same optimizing system. A major goal of the proposed work is to be able to create synthesizer systems with different voices yet the same underlying structure without a great deal of human intervention. This is extremely useful in evaluating synthesis systems independent of the voices used.

4.3 Judging the work

Judging the success of the proposed work will involve making several sets of comparisons. This work will not strictly speaking be comparable to the work of others in the field as we are not concerning ourselves with a broad range of synthetic speech problems. Rather we are addressing the narrower issue of concatenative vocabulary selection. Comparisons should be made with this limitation of scope in mind.

4.3.1 Comparing this work to others

In order to make an appraisal of the results of this work a comparison of the resulting speech to other speech will be necessary. Various versions of synthesized speech will be compared to determine if the optimization process actually improves the quality of the result. The optimizer will generate a ranking of different sets of concatenative inventories and that ranking can be checked.

the ear of the listener The ultimate arbiter of speech quality in this work will be the ear of the beholder. There are of course no arbitrary speech quality measures that are worth while for reasons that have been discussed elsewhere. The human listeners will have to judge for themselves whether the optimized elements are actually better than, equal to, or worse than the manually chosen elements.

the judgement of the audience While we will seek the broadest possible group of people to listen to and evaluate the results of this work, we will not be undertaking an extensive psychometric study of the results. Such a study is beyond the scope of the proposed work. Further, it should not be necessary to measure the results in that much detail. Future work involving fine tuning of the method might undertake a more detailed evaluation of the quality of the results.

4.3.2 Comparison to by-hand version

The most important comparison will be that between the synthesized speech created from a set of elements extracted by hand and a set of elements selected automatically. Other comparisons that could be made are between two versions of automatically produced elements and between different voices. It will not make much sense to compare the synthetic speech to natural speech or to synthetic speech from commercial systems. Speech from commercial systems incorporate a great many minor improvements in areas that are irrelevant to the proposed work. These improvements make it very difficult to compare commercial systems to experimental systems. The comparison that we will be looking to to demonstrate the success of the approach will be between synthetic speech from manually chosen elements and synthetic speech from automatically chosen elements.

possible outcomes There are three possible outcomes from the comparison between synthetic speech from automatically chosen concatenative elements and synthetic speech from manually chosen elements. The automatically chosen version could sound significantly better than, worse than, or roughly equal to the manually chosen version.

automatic better than manual If the synthetic speech generated from concatenative elements that were automatically chosen sounds significantly better than speech from manually chosen elements then we will have achieved the most desirable result possible from these experiments. In this case, a casual listening test will reveal that the automatically created elements form better transitions and work with each other better than their manually generated counterparts.

The usefulness of this result will be significant. A strongly positive result in the work will prove useful to speech research in a wide variety of ways. Some of the more useful aspects of a positive result are noted below:

- This result will allow the rapid creation of new voices for synthetic speech systems which use concatenative synthesis. Currently approximately two man years are required to create a new voice of good quality. This time would be reduced dramatically.
- This result will allow the rapid prototyping of new speech synthesis systems. Currently the vocabulary creation is a major bottleneck. In addition, the use of an automatic element generator would create much greater consistency of results than is otherwise possible.
- This result would indicate that the initial element cut points are not critical. This would suggest that a speech recognition system could be used for this initial step of element selection.
- This result would indicate that automatic methods can be expected to ultimately produce synthesizers of higher quality than has been possible with manual methods.
- This result would go a long way toward validating the speech distance measures developed in this work. These measures could be useful in a broad range of other applications.

automatic equal to manual If the synthetic speech generated from automatically selected elements sound substantially the same as speech generated from manually selected elements then our results are still quite significant. This result means that the method does at least a competent job of picking elements. All of the above results can be expected to apply with one exception. The method will probably not be expected to yield synthesizers of higher quality than manual methods can produce without further refinements.

automatic is worse than manual In the event that the speech produced from automatically selected elements is significantly worse than speech produced from manually selected elements our work will still have significance. In this case it is important to consider how the work could have this result.

- The distance measures developed could fail to give consistent results. The measures could fail to have the ability to distinguish good elements from bad. In this case, other measures could be tried and there is good reason to believe that some measure can be found that at least does not rank bad speech over good. If some sort of useful measure can not be found then the method will indeed be of no use. We consider this outcome extremely unlikely. A null result with respect to the distance measure would itself be of considerable importance as it would contradict many long standing assumptions about speech distance measures.

- The optimizer could fail to consider a wide enough variety of cases to find good elements. Since we will be starting the optimizer off with manual selections, the system will have to actively make the wrong choices in order to fail in this way. An important problem to consider here is that the optimizer has no concept of the significance of an individual word or element combination. This means that the optimizer could sacrifice the quality of a very common element combination to get slight improvements in very uncommon element combinations. This is not an insoluble problem; but it is a problem that we hope to not have to deal with in the proposed work.

References

[Allen et al. 79] J. Allen, S. Hunnicutt, R. Carlson, and B. Granstrom. Mitalk-79:the 1979 mitalk text-to-speech system. In J. J. Wolf and D. H. Klatt, editors, *ASA-50 Speech Communiction Papers*, pages 507–510, Acoustical Society of America, New York, 1979.

[Barnwell 79] Thomas P. Barnwell. Objective measures for speech quality testing. *Journal of the Acoustical Society of America*, volume 66(6):1658–1663, December 1979.

[Barnwell 80] Thomas P. Barnwell. A comparison of parametrically different objective speech quality measures using correlation analysis with subjective quality results. In *ICASSP80*, pages 710–713, 1980.

[Browman 80] Chatherine P. Browman. Rules for demisyllable synthesis using lingua, a language interpreter. In *ICASSP-80*, pages 561–564, 1980.

[Charot et al. 86] Francois Charot, Patrice Frison, and Patrice Quinton. Systolic architectures for connected speech recognition. *IEEE Transactions on Acoustics Speech and Signal Processing*, 34(4):765–779, 1986.

[Coker 76] Cecil H. Coker. A model of articulatory dynamics and control. *Proceedings of the IEEE*, 64(4):452–460, April 1976.

[Elovitz et al. 76] H.S. Elovitz, R. Johnson, A. McHugh, and J.E Shore. Letter-to-sound rules for automatic translation of english text to phonetics. *IEEE Transactions on Acoustics Speech and Signal Processing*, volume 24(6):446–473, December 1976.

[Flanagan 72] J. L. Flanagan. *Speech Analysis Synthesis and Perception.* Springer-Verlag, New York, 1972.

[Fujimura 76] Osamu Fujimura. Syllables as concatenated demisyllables and affixes. presented at the 91st meeting of the Acoustical Society of America, 1976.

[Gray and Markel 76] Augustine H. Gray and John D. Markel. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(5):380–391, October 1976.

[Hecker and Guttman 67] Michael H. L. Hecker and Newman Guttman. Survey of methods for measuring speech quality. *Journal of the Audio Engineering Society*, 15(4):400–403, October 1967.

[Itakura 75] Fumitada Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics Speech and Signal Processing*, 23(1):67–72, February 1975.

[Itakura and Saito 70] Fumitada Itakura and Shuzo Saito. A statistical method for estimation of speech spectral density and formant frequencies. *Electronics Communications of Japan*, 53-A:36–43, 1970.

[Kaplan and Lerner 85] Gadi Kaplan and Eric J. Lerner. Realism in synthetic speech. *IEEE Spectrum*, Volume 22(4):32–37, April 1985.

[Kutik et al. 83] E. Kutik, W. Cooper, and S. Boyce. Declination of fundamental frequency in speakers' production of parenthetical and main clauses. *Journal of the Acoustical Society of America*, volume 73(5):1731–1738, May 1983.

[Liberman 87] Mark Liberman. private communication, 1987.

[Lovins 78] Julie B. Lovins. A demisyllable inventory for speech synthesis. 1978. Unpublished paper, Bell Laboratories.

[Macchi 80] Marian J. Macchi. A phonetic dictionary for demisyllabic speech synthesis. In *ICASSP-80*, pages 565–567, 1980.

[Matsumoto and Wakita 79] Hiroshi Matsumoto and Hisashi Wakita. Frequency warping for nonuniform talker normalization. In *ICASSP-79*, pages 566–569, 1979.

[Olive 80] J. Olive. A scheme for concatenating units for speech synthesis. In *ICASSP-80*, pages 568–571, 1980.

[Olive and Spickenagel 76] Joseph P. Olive and N. Spickenagel. Speech resynthesis from phoneme-related parameters. *Journal of the Acoustical Society of America*, 59(4):993–996, April 1976.

[Parsons 86] Thomas W. Parsons. *Voice and Speech Processing.* McGraw Hill, New York, NY, 1986.

[Polish 86] Nathaniel Polish. A survey of unrestricted text-to-speech synthesis. May 1986. Unpublished survey paper, Columbia University Department of Computer Science.

[Rabiner and Schafer 78] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals.* Prentice-Hall, Englewood Cliffs, New Jersey, 1978.

[Roucos and Dunham 85] Salim Roucos and Mari O. Dunham. A comparison of two methods for very-low-rate speech coding. In *1985 IEEE Military Communications Conference*, pages 609–613, 1985.

[Roucos and Wilgus 85a] Salim Roucos and Alexander M. Wilgus. High quality time-scale modification for speech. In *ICASSP-85*, pages 493–496, 1985.

[Roucos and Wilgus 85b] Salim Roucos and Alexander M. Wilgus. The waveform segment vocoder: a new approach for very-low-rate speech coding. In *ICASSP-85*, pages 236–239, 1985.

[Rudnicky 81] Alexander I. Rudnicky. *CMUtalk adapted from MITtalk.* Technical Report, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA 15213, 1981. licensed Pascal source code.

[Schwartz et al. 79] R. Schwartz, J. Klovstad, J. Makhoul, D. Klatt, and V. Vue. Diphone synthesis for phonetic vocoding. In *ICASSP-79*, pages 891–894, 1979.

[Voiers 77] W. D. Voiers. Diagnostic acceptability measure for speech communications systems. In *ICASSP-77*, pages 204–207, 1977.