BOLT BERANEK AND NEWMAN INC

CONSULTING · DEVELOPMENT · RESEARCH

BBN Report No. 3508 February 1977

A FEASIBILITY STUDY OF VERY LOW RATE SPEECH COMPRESSION SYSTEMS

Final Report
Contract No. MDA 903-75-C-0180
19 July 1976 to 18 January 1977

Submitted to:

Defense Advanced Research Projects Agency
Strategic Technology Office
1400 Wilson Boulevard
Arlington, VA 22209

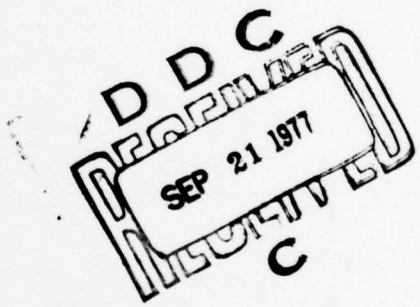
Attention: Col. Gerard R. Pepin

BOLT BERANEK AND NEWMAN INC

CONSULTING • DEVELOPMENT • RESEARCH

BBN Report No. 3508 February 1977

A FEASIBILITY STUDY OF VERY LOW RATE SPEECH COMPRESSION SYSTEMS



Final Report
Contract No. MDA 903-75-C-0180
19 July 1976 to 18 January 1977

Submitted to:

Defense Advanced Research Projects Agency
Strategic Technology Office
1400 Wilson Boulevard
Arlington, VA 22209

Attention: Col. Gerard R. Pepin

Table of Contents

1. INTRODUCTION

1.1 Overview

This final report documents our six month study into the feasibility of very low rate (VLR) speech compression systems. Much of the work was performed during the final three months following the termination of the ARPA speech understanding project. Most of the techniques used were a marriage of techniques and ideas already developed under the ARPA speech understanding and speech compression projects; others emerged during the study itself.

The feasibility study had three tasks:

(1) To design a formant-type vocoder software system operating at average rates of 500 bps (bits per second).

(2) To design a phonetic vocoder at close to 75 bps.

(3) To make recommendations for future possibilities and directions to take.

Below, we give a brief summary of the results of the study. More detailed description is given in the body of the report.

1.2 Summary of Results

1.2.1 Hybrid Formant-LPC Vocoder

The purpose of this vocoder was mainly to see how much one can stretch existing vocoder techniques in terms of lowering the

bit rate, and to compare the results with existing systems operating at fixed rates of 600 bps.

The result of our work here was a hybrid formant-LPC variable rate vocoder operating at average rates of 500 bps. The hybrid aspect of the system and the formant tracking routine were carried over from the speech understanding system, while the variable rate aspect and the quantization and coding of LPC parameters were carried over from the speech compression project. In addition, formant quantization and coding routines had to be developed. The synthesis part of the system used techniques from both projects.

In spite of the fact that this was a first attempt, the intelligibility of the resulting system was very good. However, there was some loss of naturalness and speaker identifiability. It was clear that further work would help improve the quality substantially.

1.2.2 Phonetic Vocoder

A substantial part of our effort went into developing a phonetic vocoder that would operate at average rates of 75 bps. Again, in designing this system, we used techniques from the speech understanding and compression projects.

The acoustic-phonetic part of our speech understanding system had to be modified to give a single stream of best

first-choice segments instead of a lattice of segments. The percentage of correct first choices had to be improved considerably for the phonetic vocoder application. Furthermore, an interface between the recognition and synthesis-by-rule programs was written to accommodate differences in the phonetic sets of the two system components. Considerable effort went into quantizing and encoding the pitch and duration of segments to maintain the requisite intonation of the speech while minimizing the transmitted bit rate.

The result was a phonetic vocoder that operated at average rates of 75 bps when the speaking rate was 10 phonemes per second, or an average of 7.5 bits/phoneme, which includes 2.75 bits for pitch and duration, and 4.75 bits for phonetic encoding. From the subjective listening tests we performed with a panel of listeners, we were able to make the following conclusions:

(1) The present encoding of pitch and duration at 2.75 bits/phoneme is adequate to preserve the natural intonation of the speech signal.

(2) Phonetic recognition accuracy is the principal determiner of intelligibility. A phonetic recognition accuracy of at least 80%, and preferably 85%, is necessary for high intelligibility of continuous speech in context.

(3) Synthesis is the principal determiner of speech quality at the receiver. The speech from the synthesis-by-rule program was somewhat machine-like; modifications are necessary to render the speech more natural.

1.3 Future Prognosis

The last part of this pilot project was spent in a concentrated effort to determine the realistic feasibility of actually building a real time, natural sounding 75 bps vocoder. Our conclusion was that a phonetic vocoder with

(1) 85-90% phonetic recognition,

(2) natural sounding synthesis,

(3) 75 bps average transmission rate,

(4) operating in real time,

(5) using off-the-shelf technology,

is feasible. The requisite effort is on the order of 2.5 man-years per year for a period of 3 years. The details for such an effort will be submitted in a separate document.

We emphasize that the so-called "Donald Duck" quality that is characteristic of certain synthesis programs, especially synthesis-by-rule programs, is not a necessary thing. Synthesis of natural sounding speech is definitely possible with a relatively small effort. The major development has to be in storing many of the speaker characteristics, instead of synthesizing them by rule.

1.4 Report Outline

Section 2 gives an overall description of a speech compression system. Details of the 500 bps hybrid formant-LPC

vocoder are given in Section 3. Section 4 contains the description of the 75 bps phonetic vocoder, along with recognition and listener performance results. In Sections 5 and 6 we present our conclusions regarding very low rate speech transmission and give our recommendations for a feasible very low rate phonetic vocoder.

2. SPEECH COMPRESSION

2.1 Components of a Speech Compression System

Figure 1 shows the various components of a speech compression system. The first component analyzes the speech signal s(t) that has been low-pass filtered and time-sampled, and extracts a vector of unquantized parameters $\underline{x}(t)$. These parameters are then quantized and encoded in the encoder as $\underline{y}(t)$ and are transmitted through the transmission channel. In a noiseless channel $\underline{y}'(t)=\underline{y}(t)$. (This is generally the case, for example, in the ARPA Network.) The parameters $\underline{y}'(t)$ are decoded in the decoder to produce an estimate $\underline{x}'(t)$ of the analysis parameters $\underline{x}(t)$. The last component in Fig. 1 uses the parameters $\underline{x}'(t)$ to synthesize the signal s'(t) which is an approximation to the original signal s(t). The nature of the synthesizer (model) dictates the type of analysis to be performed. Figure 2 depicts the two major components of the synthesizer: excitation and transfer function.

2.2 Fidelity Criterion

In any communication system, one must specify a fidelity criterion that is optimized in the system design. Rate distortion theory (see, for example, [1]) has provided a mathematical basis for many types of data compression. However, the theory requires that one specify the distortion function to

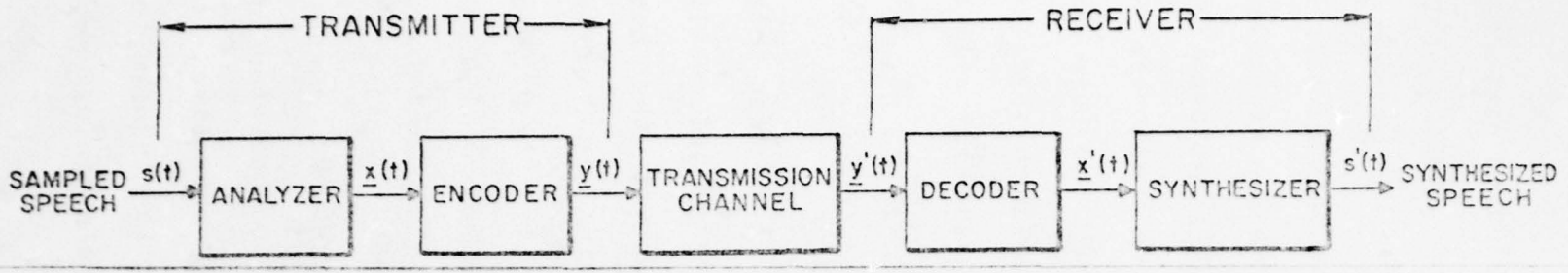


Fig. 1. Components of a speech compression system

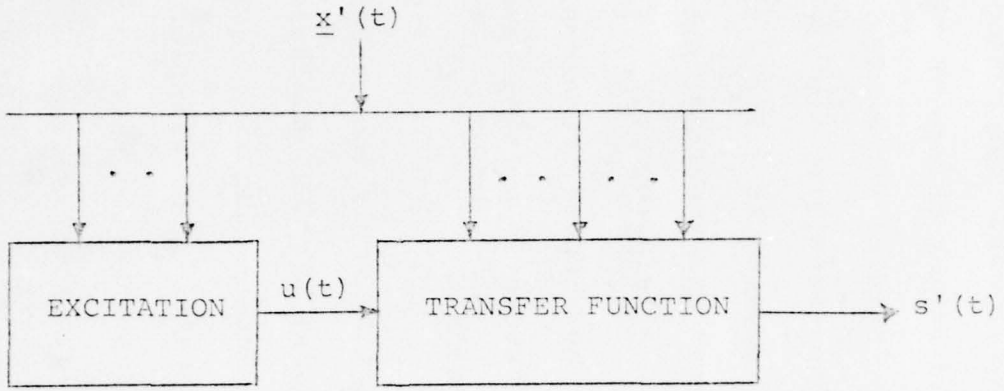


Fig. 2. Major components of a speech synthesizer.

be minimized. Using some of the generally used distortion functions (such as mean square error) has been of limited value in speech compression. The reason is that the "true" distortion function should be the change in "quality" of the speech signal, and the mean square error is not a good measure of quality reduction. The inability to devise completely mathematical (objective) equivalents for the subjective criteria of "intelligibility" and "quality" makes it difficult to use the results of rate distortion theory to the fullest. Such objective criteria are now under research [2,3], and hopefully will be useful in the future in putting speech compression on a more mathematical basis.

This is not to imply that speech researchers have been idle. To the contrary, much use has been made of experimental results on the quality determining factors in speech. One of the most important findings, for example, has been that the ear is relatively insensitive to phase; a fact that has been used extensively in designing different types of vocoders.

For our purposes here we shall use the word "quality" to include such aspects of the signal as naturalness and speaker identifiability; and "intelligibility" to refer to understanding of the message, irrespective of the quality. Our objective, in designing a speech compression system, then, is to minimize the number of bits/second in $\underline{y}(t)$ in Fig. 1, while maintaining good

quality and intelligibility in the synthesized speech signal s'(t).

3. HYBRID FORMANT-LPC VOCODER

3.1 Introduction

As part of our low-rate systems development, we have developed a hybrid formant-LPC vocoder capable of transmitting intelligible speech at 500 bps. Our system transmits at a variable rate, in contrast with the fixed-rate 600 bps system developed by Kang at NRL [4]. In a variable rate system, we transmit the speech parameters asynchronously and only when they differ sufficiently from the last transmitted values. The three parameters being transmitted are pitch, gain (energy) and a set of parameters representing the spectral shape.

We call our vocoder hybrid because it combines some features of a formant vocoder with those of a linear predictive type. These features have to do with the spectral models used to define the parameters of the waveform synthesizer. We use two alternative spectral models, one being the linear prediction spectrum defined by the optimal 4-pole model. The other is a formant spectrum defined by the first 5 formants and their bandwidths. This corresponds to a 10 pole model as each of the formant/bandwidth combinations represents a complex-conjugate pole pair.

3.2 Analysis

We sketch briefly the operation of our vocoder system beginning with the initial analysis. Given the speech wave digitized at 10 kHz, we do a digital preemphasis at zero Hz. We then compute a 13-pole model every 20 ms. This model serves as the basis for both spectral models used in the vocoder. For unvoiced speech, we choose the first 4 reflection coefficients to represent the spectrum. For voiced speech, we first solve for the roots of the polynomial defined by the 13-pole model. From the complex-conjugate pole pairs, we extract the first 3 formants with high reliability, using a formant tracking routine developed under the speech understanding project [5]. The vocoder transmits one of these two spectral models depending on the voiced/unvoiced decision. Pitch and gain are also extracted every 20 ms. We use the pitch value to make the voiced/unvoiced decision.

3.3 Transmission

As stated in the first paragraph, we transmit pitch, gain and spectral shape at a variable rate. To determine when to transmit each of these, we use a double threshold algorithm. This allows us to transmit more often during regions of rapid change. The algorithm was developed under the speech compression project; details of the algorithm are given in [6-9]. Using a thresholding algorithm requires a distance measure to compare two

values of a parameter. We use the minimum prediction residual [10] to compute the distance between spectra. For pitch and gain, we take the absolute differences of their encoded values.

Our vocoder has a basic frame rate of 50 frames per second. For timing purposes, we transmit every 20 ms a 3-bit header specifying which of the speech parameters are to be transmitted during that frame. A value of 1 in each of these three bit positions signifies transmission of pitch, gain and spectral shape, respectively. Since the choice of the spectral model to be transmitted is based on the voiced/unvoiced decision, the model is completely specified by the pitch value. This means we need not transmit an extra bit for that purpose.

The different parameters are quantized as follows:

FORMANT FREQUENCIES

	Range(Hz)	# of levels
F1	180-780	12
F2	750-2150	12
F3	1690-2890	8

The three formant frequencies require a total of 10.17 bits/frame.

LOG-AREA-RATIOS

	Range	# of levels
LAR1	(-2.5)-(12.5)	15
LAR2	(-2.5)-(8.5)	11
LAR3	(-3.5)-(4.5)	8
LAR4	(-1.5)-(4.5)	6

The four log-area ratios require a total of 12.95 bits/frame.

PITCH

Range(Hz)	# of levels
50-450	32

Pitch is quantized linearly on a log scale at 5 bits.

GAIN

Range(dB)	# of levels
0-40	16

Gain requires 4 bits of quantization.

Assuming pitch, gain and spectral shape are all transmitted in the same frame, we have rates of 19.17 and 21.95 bits per frame for voiced and unvoiced frames respectively. The transmission thresholds were adjusted such that the spectral

shape was updated 25-30 times per second while pitch and gain were updated 15-20 times per second. (Figure 3 shows the spectral transmission points for one sentence.) These rates produce an overall transmission rate of 500-600 bps. With entropy coding [11], the average transmission rate goes down to 500 bps or less.

3.4 Synthesis

To resynthesize the speech wave, we use a direct-form synthesis structure because of its relative insensitivity to errors in formant tracking. The synthesizer operates at a basic frame rate of 10 ms, requiring the decoded speech parameters to be linearly interpolated. Note that the decoded pitch value is not interpolated across voiced/unvoiced boundaries.

In constructing the synthesis filter, the system checks to see if that frame is voiced. If so, it approximates the formant bandwidths based on their target frequencies. Our approximations are as follows:

B1= 60 Hz

B2= 50 + F2/80 Hz

B3= 130 Hz.

In addition, we use fixed 4th and 5th formants to add crispness to the voiced speech. Their values are:

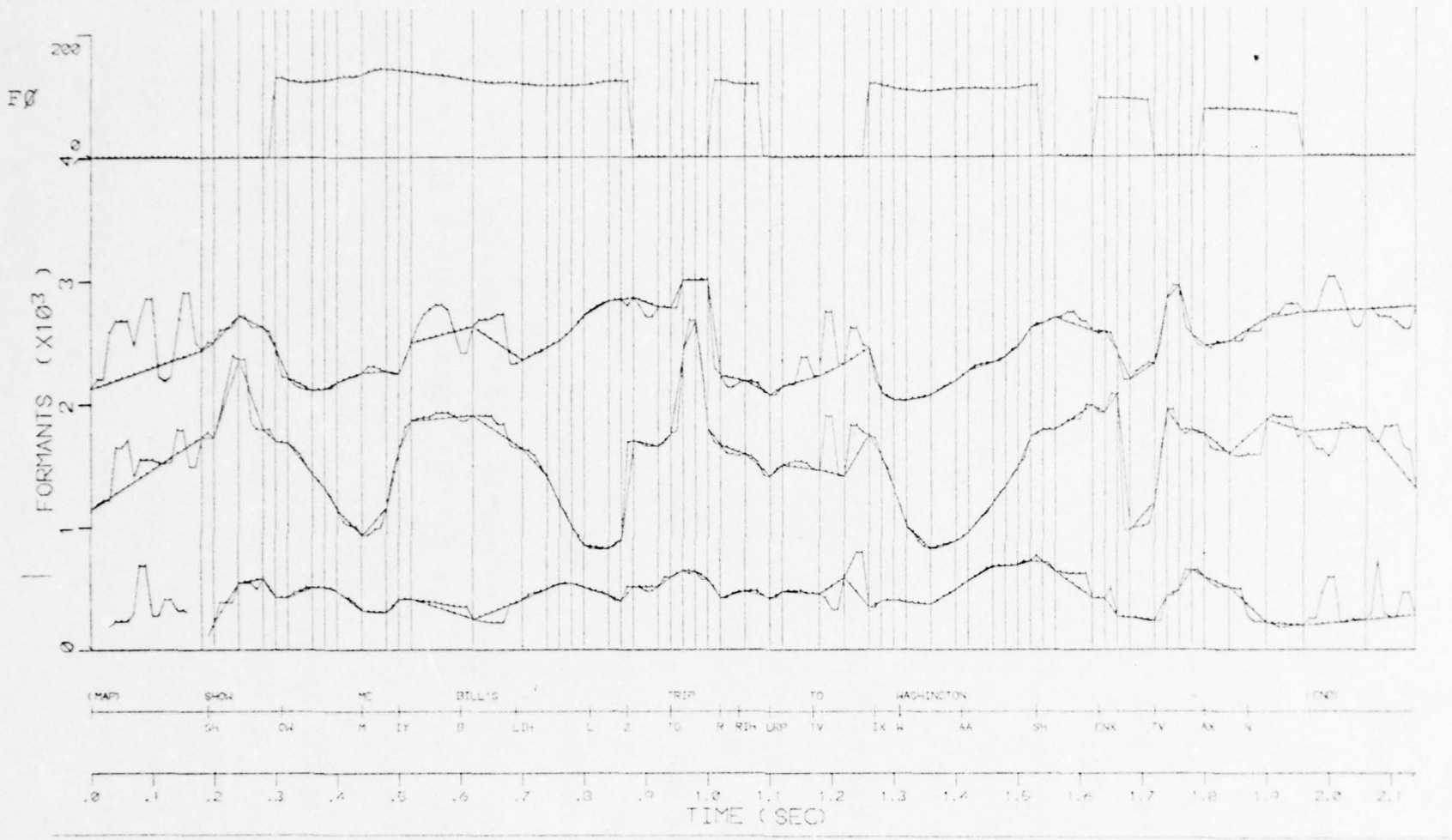


Fig. 3. Variable rate transmission. Vertical lines indicate time points at which spectral parameters are transmitted. The first three formants and pitch are shown, along with the interpolated formants between transmission points.

F4= 3250 Hz, B4= 200 Hz

F5= 3700 Hz, B5= 250 Hz.

If a frame is unvoiced, the 4 transmitted log-area ratios are converted to predictor parameters [12], after appropriate deemphasis. For both voiced and unvoiced spectra, we add a deemphasis filter (single real pole) to restore the spectral tilt. This results in the voiced and unvoiced synthesis filters having 11 and 5 poles respectively. In both cases, we assume a 10 kHz sampling rate.

3.5 Testing and Conclusions

Our set of test utterances consisted of six utterances from a single speaker. Our vocoder system was tuned to this speaker only insofar as the formant frequency ranges were determined from a large number of sentences from this speaker. These ranges would probably suffice for any male speaker.

The system requires approximately 15-20 times real time to process an utterance on our PDP-10. This includes all of the variable rate computations, encoding, decoding and synthesis.

The intelligibility of the synthetic speech was so high that we did not feel it was necessary to perform a formal test in this short-term project. However, some speech naturalness was lost, along with speaker identifiability. The voicing problems normally associated with vocoders were somewhat exaggerated in

this system because of the dependence of the spectral shape (whether 11 or 5 pole) on the voicing decision.

With further work, we are confident that the bit rate can be reduced further, with an increase in quality. Quantization and encoding can be optimized, especially for the log-area ratios. Using our recently developed techniques from the speech compression project, we can make better transmission decisions, thus lowering the transmission rate or increasing the quality [13]. Our rudimentary choice of formant bandwidths can be substantially improved to result in more natural synthetic speech.

Overall, we feel that 500 bps is a viable rate at which intelligible and reasonably good quality speech can be transmitted. Transmission at these rates has important applications in underwater and underground speech communications.

4. PHONETIC VOCODER

4.1 Introduction

In order to design a speech compression system that transmits at 100 bps or less, one has to deal with speech segments on the order of a single sound or phoneme. If shorter segments were used, the number of bits needed would be too large. In addition to the identity of the speech sounds, stress and intonation information is essential to the naturalness of the speech, and to the semantic information contained in the inflections. Such information is carried in the speech signal via the pitch and duration of the segments, and to a much lesser extent by the signal amplitude. Therefore, in addition to transmitting the identity of the sounds, it is also helpful to be able to transmit pitch and duration. Proper pitch and duration information also helps to mask some of the inevitable phonetic errors made in the recognition of the individual phonemes. In our phonetic vocoder, we transmit phonetic as well as pitch and duration information at a total average data rate of 75 bps.

There are two major obstacles in the design of a phonetic vocoder: recognition and synthesis. The recognition component attempts to identify the particular set of phonemes that were uttered by the speaker. Any error in the recognition process results in the transmission and final synthesis of the wrong phoneme. Since the phoneme is the smallest unit of meaning in

the language (i.e. two words differing by only one phoneme usually have different meanings, e.g., pad and bad), every phonetic error is likely to cause a problem in intelligibility. In a phonetic vocoder, therefore, recognition is the principal determiner of intelligibility. Fortunately, in continuous speech, context provides much redundancy and a good number of phonetic errors are often corrected for unconsciously by the listener.

Given a sequence of phonemes, the synthesis component attempts to resynthesize the intended message. Most phonetic synthesizers, when given the correct sequence of phonemes, produce speech with good intelligibility. Therefore, if intelligibility were our only concern, most any type of synthesis would do. However, for communication purposes, one is also concerned with the quality and naturalness of the speech. To that extent, the use of transmitted pitch and duration is helpful, but not sufficient; the phonetic synthesis must also produce natural sounding speech. In summary, synthesis is the principal determiner of quality.

Below, we present a brief description of our phonetic vocoder, followed by the results of listener judgments on the intelligibility of the synthesized speech.

4.2 Overall Vocoder Description

Beginning with existing programs from the speech understanding and compression projects, we implemented a phonetic vocoder on our PDP-10 system. Figure 4 shows a block diagram of the whole process. The speech signal is sampled at 10 kHz and processed by our signal analysis program, which produces time-varying parameters such as energy, formants, zero-crossings, etc. Then the Acoustic-Phonetic-Recognition (APR) program [14,15] from our speech understanding project uses these parameters to produce a phonetic transcription in the form of a segment lattice with associated probabilistic phoneme scores. The phonetic transcription is then examined by the Interface program (described in Section 4.5) which finds a single best path transcription that can be quantized and coded and sent to the synthesis-by-rule program (see Section 4.4). This program takes the phonemes and their associated duration and pitch, and applies several rules that result in parameter tracks that are then used to generate the output synthesized speech signal. The major components shown in Fig. 4 are described further below.

4.3 Acoustic-Phonetic Recognition (APR)

In the APR component of our Speech Understanding System (SUS) we try to score many possible alternative phonetic transcriptions. These transcriptions differ both in the number of phonetic segments and in the phoneme scores in each segment.

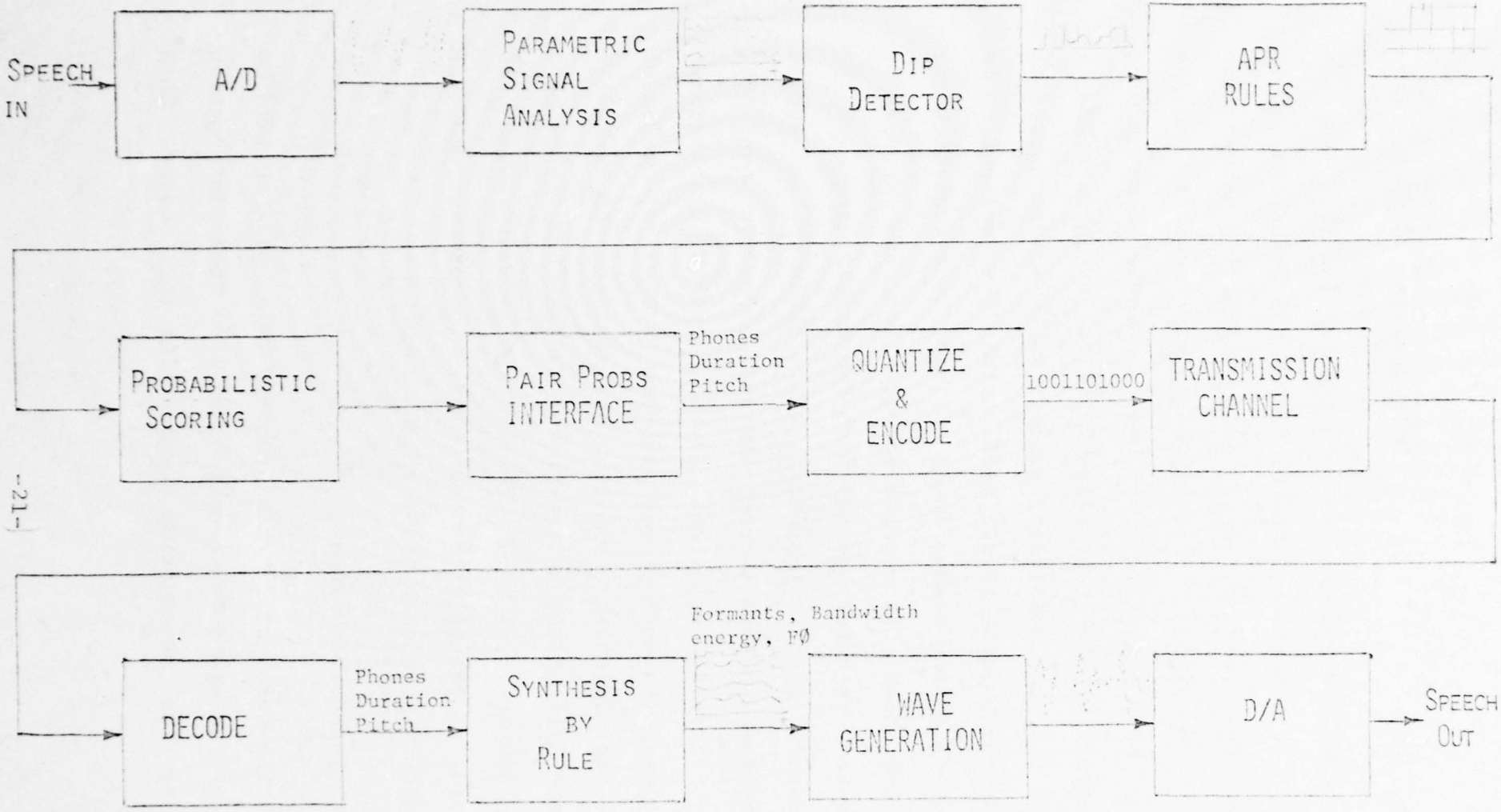


Fig. 4. Block diagram for phonetic vocoder. It utilizes Acoustic-Phonetic Recognition and Synthesis-by-Rule programs.

We accomplish this through the use of a segment lattice. Fig. 5 shows a section of a segment lattice. There are several possible paths through the lattice, where each segment (between two vertical lines) corresponds to a single phonetic segment. Since we cannot be absolutely sure of the phonetic identity of a segment, we associate with each segment a list of probabilistic scores - one for each possible phoneme. These scores depend on acoustic measurements made in the vicinity of the segment. In our speech understanding system it is not very important that the correct phoneme have the highest score, since other sources of knowledge are used to determine the most reasonable sequence of words. However, it is important that there be a strong correlation between phoneme correctness and score. In our SUS, the lexical retrieval component gives scores to hypothesized words by combining the appropriate phoneme scores on adjacent segments in the lattice. Finally, the semantic and syntactic components determine the most likely sentence that spans the hypothesized set of words [16]. These "higher-level" components are absent in a phonetic vocoder and, therefore, phonetic recognition accuracy becomes of paramount importance.

4.4 Speech Synthesis-by-Rule

Through the use of a synthesis-by-rule program [17], a synthetic speech waveform is reconstructed from the information transmitted by the recognition component of the vocoder. This

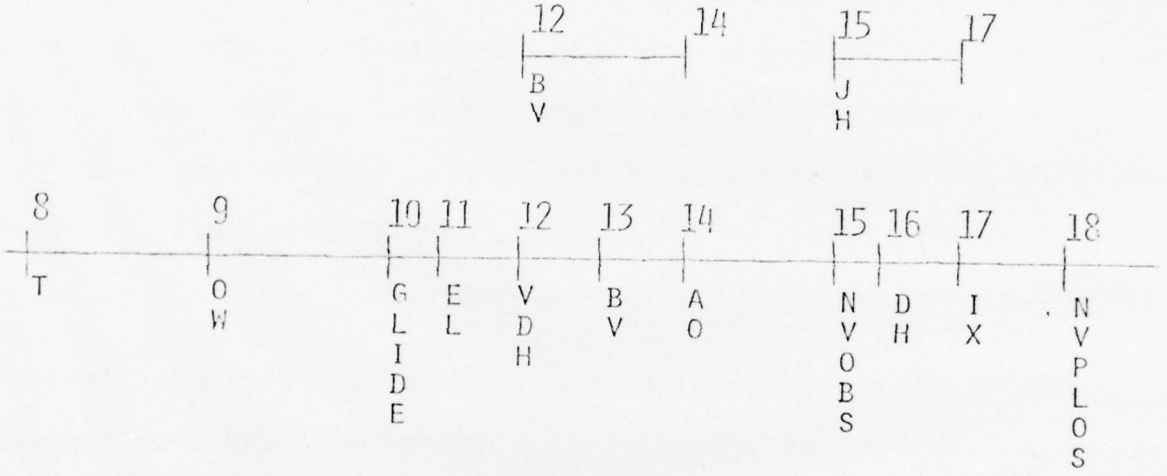


Fig. 5. Section of a segment lattice representing the words "total budget".

program was originally designed for use in our speech understanding system as an audio response generator and as a spectral pattern generator for word verification. In that application, the input was an abstract phonemic and syntactic representation of each sentence to be spoken. The phonological component of the program transformed this representation into a narrow phonetic transcription and a specification of a duration and fundamental frequency target (pitch) for each phone. The acoustic-phonetic component of the program then derived a set of time functions (formant frequencies, source amplitudes, etc.) to control a terminal analog speech synthesizer.

The synthesis-by-rule program was modified for use in the phonetic vocoder by removing the above-mentioned phonological component and by translating the vocoder output phone inventory into the phonetic inventory of the rule program. Other synthesis work performed during this period concerned improvements to the synthesis quality by comparing detailed spectra of selected synthetic speech sounds with comparable spectra obtained from the same utterances spoken by an adult male talker.

4.5 Interface

In a phonetic vocoder it is not meaningful to have several phonemes with scores on competing paths through a lattice. Therefore one must pick a single "best path" through the segment

lattice, using the most likely phonemes at each point. This "best path" transcription can be encoded along with measured duration and pitch and then be sent to the Synthesis-by-Rule (SBR) program to be decoded and re-synthesized back into speech.

The APR program and the SBR program operated independently in the SUS. There were several inconsistencies between them that needed to be taken into account. One basic difference was in the phoneme sets used by the two programs; the APR program had 71 phonemes, while the SBR program had 50 different phonemes. In some cases the differences could be accounted for by a simple mapping. (For example, the ARPABET symbol NX must be translated into the SBR symbol NG.) Other changes required some contextual information. For example, if the phoneme /l/ is preceded by a vowel and followed by a consonant (i.e., it is "postvocalic"), it must be called "LX" instead of "L" so that the SBR program knows to synthesize it differently from the prevocalic case. Since postvocalic /r/ strongly affects the preceding vowel, a phoneme sequence, such as /i-r/, is replaced by a single segment, labeled "IR", which is treated specially.

There are also several differences in the definitions of phoneme boundaries. For example, an unvoiced plosive (such as P,T,K) which is followed by a vowel or sonorant results in three distinct acoustic regions. First there is a period of silence corresponding to total vocal tract closure. Then there is a

short burst of frication due to air turbulence at the release. This is followed by a period of aspiration ("h-like" noise) before voicing starts during the vowel or sonorant. In the APR, all three periods are considered to be part of the unvoiced plosive, since they each contain some clues as to the identity of the plosive. In the SBR program, the aspiration is considered part of the following vowel or sonorant, since the formants are excited during this time.

A program has been developed as a test bed for several possible strategies. One phase in this program applies several transformation rules to convert a transcription from the APR conventions to the SBR conventions.

The interface program has three basic methods for determining the "best path" through a segment lattice. First, as a control, the program can start with a "hand labeled" (i.e., correct) phonetic transcription of a sentence using APR conventions. This allows us to determine the effect of parameter quantization and the loss in naturalness due to the SBR program. In the second (intermediate) method, the program starts with the segment lattice produced by the APR, but is given the correct path of segments through the lattice. In this case, the program must still decide on the phonemes to be transmitted (the major difficulty). The third method requires that the program choose what seems to be the best segmentation path, and also choose the phonemes along that path.

The program also allows the user to determine the type and degree of quantization of pitch and duration to be used in order that we could examine the effect of different amounts of quantization.

4.6 Recognition Performance

In order to try out several different schemes for determining the "best path" phonetic transcription from the segment lattice, we needed a data base of sentences. We chose to use the official test set of 132 utterances used in testing our SUS. These 132 utterances represented the speech of 3 speakers, and had not been used at all in training the APR program. We also chose a subset of 13 of these sentences that were spoken by one speaker (WAW).

Given a sequence of segments through a segment lattice, as in the second method, there are several possible schemes for choosing the phonemes. Since each phoneme has been assigned a score, the simplest method is to pick the phoneme on each segment which was given the highest acoustic score by the APR program. Figure 6 shows the percentage of phonemes that were correct on the first choice (out of 71 possible phonemes) for each of the two data bases. As can be seen from the first line, the 13 WAW sentences performed slightly better (58% vs 54%) than the whole test set, but there was not a major difference since the program was not tuned specifically to that one speaker. In order to

No. of sentences and (speakers)	132 (3)	13 (1)
Highest scoring phoneme	54%	58%
+ a priori phoneme probability	57%	61%
+ a priori diphone probability	—	77%

Fig. 6. Phonetic recognition performance, given the correct path.

assess the intelligibility of synthetic speech when the percentage of correct phonemes was only 58%, we synthesized all 13 sentences using the SBR program and played them to a group of listeners. The intelligibility was so low that it was clear to us that the recognition accuracy must increase substantially before we could have any meaningful test of a potentially realistic phonetic vocoder. So, our next effort was to improve recognition performance.

Our first attempt was to include in the acoustic score for each phoneme its a priori probability of occurrence, and then choose the phoneme with the highest score. As can be seen from Fig. 6, the recognition performance increased by only 3%. Thus, our first attempt did not prove to be very useful. We should stress here that our intention was to increase the performance of the best first choice task with as little effort as possible, due to the short-term aspect of the project. Thus, actually changing the recognition algorithm was out of the question because it would have been a very involved task.

Up to this point, each phoneme decision had been independent of adjacent phoneme decisions. However, it is known that some phoneme sequences are much more likely in English than others. Using our data base, we determined the a priori probability of each possible pair of phonemes. Now, for each phoneme on a given segment, we combined the acoustic score on that phoneme with the

a priori probability of that phoneme occurring in a pair with each of the possible 71 phonemes on the preceding segment. Then, using an iterative procedure, we found the sequence of phonemes that maximizes these combined scores, and results in a consistent phoneme string. Using this method, the percentage of correct phonemes increased to 77% for the 13 sentences. (The experiment was not run for all 132 sentences in this case.) Upon informally listening to synthesized versions of the 13 sentences, it was clear that some were very intelligible and others were not. It turned out that while the average recognition accuracy was 77% over all the sentences, some had substantially better performance and others worse. Since one of our objectives was to determine the level of performance that would be meaningful for a communications task, we decided to choose those sentences that had better performance for further formal testing. The results of the test are given in Section 4.8.

4.7 Quantization and Encoding

4.7.1 Phoneme Encoding

In order to minimize the number of bits of information needed to transmit the phonemes, we must code them efficiently. Since we define 71 different phonemes, this would require a 7 bit code for each phoneme. However, since we had determined that some pairs of phonemes are much more likely than other pairs, we decided to use entropy coding [11] to reduce the data rate. As

it turned out, the average number of bits needed to code a pair of phonemes using entropy coding was 9.5 bits. For an average speaking rate of 10 phonemes/second (or 5 phoneme pairs/second), this method then resulted in an average transmission rate of 47.5 bps to encode the phonemes.

4.7.2 Duration

Phoneme durations vary between 20 ms and 300 ms. A difference in duration of more than 10 ms is perceptible for most phonemes, especially the short ones. This would imply a need for 10-15 different choices of phoneme duration. However, some phonemes are consistently longer than others. Also, phoneme duration is perceptually more important for some phonemes than for others. Another important fact is that phonemes in the syllable preceding a pause or silence are lengthened significantly. With these three facts in mind we designed the program to accept a different set of allowable durations for each of the 71 phonemes. Thus the phoneme [IY] was allowed to be one of (55 75 95 110 140) ms, while the phoneme [V] only needed to be quantized to one of (30 50 70) ms. In addition, associated with each phoneme is a lengthening factor to be used to account for prepausal lengthening. Figure 7 shows, for each phoneme, the lengthening factor, followed by the allowable durations. For example, since the lengthening factor for phoneme [IY] is 1.5, then the encoded duration at the end of a phrase will be one of (82 112 142 165 210) ms.

Phoneme	Lengthening Factor	Durations
IY	1.5	55 75 95 110 140
IH	1.5	45 65 85 110
EH	1.5	60 80 100 140
AE	1.5	65 85 105 130 170
AA	1.5	70 95 125 165 200
AH	1.5	50 70 90 110
AO	1.2	65 85 110 140 180
UH	1.5	55 75 95 105
UW	1.5	65 75 105 135
AX	1.5	30 50 70
ER	1.5	60 95 110 135
EL	1.5	55 70 90 110
EY	2.0	70 100 125 150 175
OW	1.5	55 85 110 130
AW	1.5	70 95 120 145
AY	2.0	105 125 150 190 250
OY	1.5	40 60 80 120 160 220 280 350
L	1.5	30 45 65 90
W	1.5	20 30 40 60 80 110 140 180
R	1.5	30 45 60 75
Y	1.5	20 30 40 60 80 110 140 180
N	1.3	25 40 55 80
M	1.5	30 45 60 75
NX	1.5	35 50 65 85
P	1.3	60 85 100
T	1.5	20 30 40 60 80 110 140 180
K	1.5	20 30 40 60 80 110 140 180
B	1.5	50 70 85
D	1.5	30 50 70
G	1.5	45 65 80
CH	1.5	80 105 130
JH	2.0	70 85 110
F	1.5	50 80 100 120
TH	1.5	70 90 120 150
S	2.0	40 60 80 100 120
SH	1.5	80 100 120 140
V	2.5	30 50 70
DH	1.5	30 50 70
Z	2.0	40 60 80 100
ZH	1.5	40 60 80
HH	1.5	40 60 85
DX	1.5	30 45
EN	1.5	40 60 80
EM	1.5	80 100
ENX	1.5	70 90
URP	1.5	50 70 90 120
URT	1.5	40 70

Fig. 7. (Continued on next page.)

Phoneme	Lengthening Factor	Durations
URK	1.5	35 55
URB	1.5	50 70 90 120
URD	1.7	35 50 65
URG	1.7	35 50 65
IX	1.5	30 50 70
AXR	1.5	45 65 90 115
UY	1.5	40 60 80 120 160 220 280 350
EA	1.5	40 60 80 120 160 220 280 350
UAX	1.5	20 30 40 60 80 110 140 180
TX	1.5	20 30
Q	1.5	20 30 40 60 80 110 140 180
-	1.5	40 60 80 120 160 220 280 350
UIX	1.5	20 30 40 60 80 110 140 180
ST	1.5	60 80
TV	1.5	50 70 95 115
TG	1.5	80 105 120 135
TS	1.2	50 65
LIH	1.5	65 85 100 120
RIH	1.5	50 65 80
INX	1.5	50 70 95
YN	1.0	40 60 80
YM	1.0	50 65
KA	1.5	20 30 40 60 80 110 140 180
TCH	1.5	55 70 90

Fig. 7. Lengthening factor and allowable durations.

The durations to use were arrived at through use of our Acoustic-Phonetic Experiment Facility [14] and the complete data base used for our SUS. The facility was used to gather a distribution of the different durations for each phoneme. Figure 8 shows the cumulative distributions of the duration of the phoneme [IY]. The dotted line indicates the duration of [IY] divided by 1.5 for those cases when it is in the syllable preceding a silence. The vertical lines indicate the durations chosen. Dividing the prepausal data by 1.5 has brought the two curves very close to each other, so that using the five values marked ensures that the duration is rarely off by more than 5-10 ms. From these plots and a knowledge of the perceptual importance of varying duration for each phoneme, the sets of durations were chosen.

In quantizing phoneme duration, there are two requirements to be met. First, the quantized phoneme duration should be close to the measured phoneme duration. Second, the total accumulated duration must be close to the total measured duration, in order that the vocoder output keep in step with the input. In our system, if the measured phonetic duration is exactly equal to one of the allowed durations for that phoneme, that value is used. If the measured duration is between two allowed durations, then the duration that will more closely approximate the total elapsed time is used.

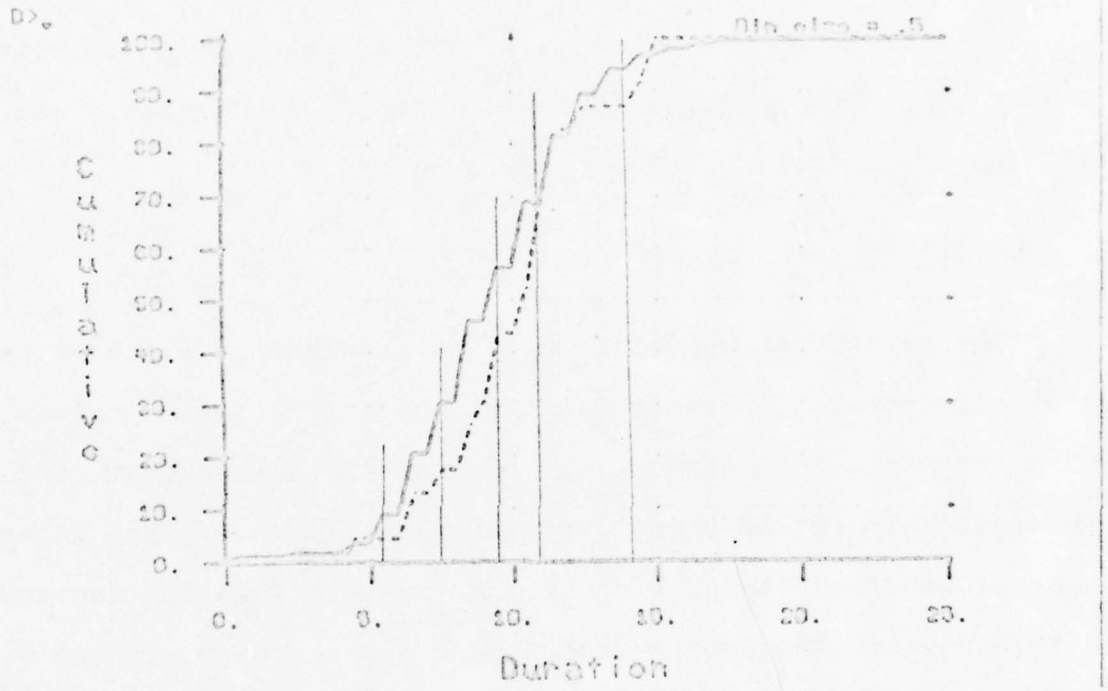


Fig. 8. Cumulative distribution of the duration of [IY].

We have experimented with different sets of durations, and have found that allowing between three and five durations for most phonemes results in speech that sounds almost as natural as when the phoneme durations are not quantized. With entropy coding, we have found that, on the average, only 1.5 bits are needed to code each segment, for a total of 15 bps, assuming average transmission of 10 phonemes/second.

4.7.3 Pitch

The synthesis-by-rule program requires a pitch value for each segment sent. It assumes that the pitch value given for each segment is reached at the end of that segment, and then connects this set of points with straight lines. The interface program handles pitch differently for vowels than for non-vowels. Within vowels, the program computes a least squares linear fit to the pitch over the vowel segment. That vowel is assigned the pitch value derived by extrapolating the line to the end of the segment. If the previous segment is not a vowel, then it is assigned a value by extrapolating the line to the beginning of the vowel. If the preceding segment is also a vowel, then it is assigned a value between those predicted by evaluating the pitch during the two segments. Consonants that are not followed immediately by vowels are assigned the pitch of the previous vowel. (Since this case can be detected from the phoneme sequence, no pitch information need be sent for these segments.)

Rather than transmitting codes specifying the measured pitch during a segment, we transmit codes reflecting changes in pitch from the last pitch value sent. This is because pitch usually changes very slowly, though it may take on a wide range of values during a sentence. If we use the five pitch differences (-30, -13, -4, +3, +16 Hz), and assume that each utterance starts with a default pitch of 130 Hz, we find that the resulting pitch closely resembles the original. Since the smaller changes are much more likely, entropy coding of these five value codes only requires an average of 1.9 bits per segment. However, since approximately one third of the segments are consonants followed by another consonant that need no pitch code, the actual average rate reduces to 1.25 bits/phoneme, or a total of 12.5 bps.

The total average information rate needed to transmit each phoneme is, therefore: 4.75 + 1.5 + 1.25 = 7.5 bits. This corresponds to a total of 75 bps for 10 phonemes/second transmission.

4.7.4 Perception of Quantized Pitch and Duration

As we experimented with different quantization schemes we compared the "quantized speech" with the "unquantized speech" in informal listening tests. The major interesting result was that even with our heavy quantization of pitch and duration, the effect on perception at the 'synthesizer was negligible. This unexpected result meant that we could afford to transmit this

information using very few bits, and simultaneously preserve most of the intonation of the original speech.

4.8 Results of Intelligibility Experiments

Eight of the thirteen sentences in the data base were grouped into a dialogue with nine other sentences (shown in Fig. 9). The numbered sentences were the output of the phonetic vocoder, and the rest of the sentences were spoken naturally. A group of nine subjects were played the dialogue, without being told anything about the subject matter of the dialogue. After each vocoded sentence was played, the tape was stopped, and the subjects were given time to write down as many words of the vocoded sentence as they could remember. When all were ready, the next pair of sentences were played. When the dialogue was finished, the experiment was redone. This time, the subjects were asked to fill in (on a different line) any words they may have missed the first time through. The results of the experiment are tabulated in Figs. 10-12.

Fig. 10 shows, for each subject listener, and for each vocoded sentence, the percentage of words written down correctly. The first and second lines in each case give the results for the first and second trials, respectively. Averages over subject, over sentence, and total averages are also shown. As can be seen, there is a wide variation across subjects and across sentences. Only subject BT, and less so LC, had been

This is HWIM. Would you like to discuss anything about travel budgets or meetings?

1) Where is the next ASA meeting?

The next ASA meeting will be in El Paso, Texas.

2) Show me all the trips to El Paso.

Jerry Wolf and Rich Schwartz are scheduled to attend the ASA meeting there.

3) How many trips has Rich taken?

Rich has taken one trip to Washington with Bill Woods, and one to Philadelphia for the IEEE Conference.

4) Show me Bill's trip to Washington.

Bill Woods has taken one trip to Washington and has two other trips scheduled.

5) How many trips are there?

Bill has taken nine trips and has two trips remaining.

6) Give me a list of the untaken trips.

Bill is scheduled to go to Washington in February to attend the ACL meeting and in March to talk with Bob Kahn about the speech budget.

7) What trips remain in the speech budget?

Jerry and Rich's trip to El Paso and Bill's two trips to Washington.

8) Are we over the budget?

No. There is $400 left.

Fig. 9. Dialogue used for intelligibility test. The vocoded sentences are numbered from 1 to 8.

Sent	SUBJECT									Avg
	DR	MB	AR	FU	BH	JB	MA	LC	BT	
1)	89	100	67	100	100	17	100	100	100	86
	94	100	100	100	100	83	100	100	100	97
2)	100	100	100	57	77	29	100	100	86	83
	100	100	100	86	100	100	100	100	100	98
3)	83	67	100	100	100	100	100	100	100	94
	83	100	100	100	100	100	100	100	100	98
4)	83	83	50	50	83	58	83	100	50	71
	91	100	83	67	100	58	67	100	50	80
5)	100	100	100	100	100	100	100	100	100	100
	100	100	100	100	100	100	100	100	100	100
6)	100	100	50	50	100	44	100	100	100	83
	100	100	50	50	100	100	100	100	100	89
7)	100	86	43	100	100	100	93	100	100	91
	100	100	57	100	100	100	93	100	100	94
8)	100	100	100	80	100	100	100	100	100	98
	100	100	100	100	100	100	100	100	100	100
Avg	95	92	74	78	95	66	97	100	92	88
	96	100	84	86	100	93	95	100	94	94

Fig. 10. Percent correct word recognition. Results of second trial are shown immediately below first trial results. Averages are given for each speaker and for each sentence.

Sent	# of words	% phoneme correct	% word correct	% sentence correct
1)	6	90	86	67
			97	78
2)	7	71	83	56
			98	89
3)	6	86	94	78
			98	89
4)	6	67	71	11
			80	33
5)	5	77	100	100
			100	100
6)	8	86	83	67
			89	78
7)	7	88	91	78
			94	89
8)	5	100	98	89
			100	100
Avg	6	84	88	69
			94	82

Fig. 11. Effect of Sentence Length and Phoneme Correctness on Intelligibility

Subjects with experience are: AR, BH, MA, LC.
Subjects without are: DR, MB, FU, JB, BT.

	Trial	w/o Exp	w/Exp	All
% Correct Words	1	83	92	88
	2	94	95	94
% Correct Sentences	1	59	78	69
	2	75	87	82

Fig. 12. Effect of listener experience on word recognition.

peripherally associated with the SUS project, and therefore had some idea of the subject matter in the sentences. Note that the performance of BT was not much better than the average. LC had the additional advantage that she had had experience listening to vocoded speech in general (see below).

It is clear from Fig. 10 that the results for the second trial were almost always better than the first trial. Other than simple repetition, there are two main reasons why the performance was better in the second trial. First, after the first trial, the subjects became familiar with the subject matter of the dialogue and, therefore, were able to use context more fully to disambiguate words. Second, the subjects became familiar with speech produced by a SBR synthesizer, and hence were able to understand it better in the second trial. The latter point will be discussed further below.

From the results of the experiment, it was clear that the most important factor that determines the intelligibility of the sentences is the accuracy of the phonetic recognition of the vocoder. Another important factor turned out to be the length of the sentence; shorter sentences were generally easier to understand. Fig. 11 shows the effect of these two dimensions on the intelligibility of the eight sentences used. The number of words in each sentence is shown in the first column, and the recognition accuracy is shown in the second column. Note that

recognition accuracy is an objective measure of the performance of the recognition component of the vocoder; the accuracy is defined as the percentage of phonemes that were recognized correctly by the recognition component. The third and fourth columns in Fig. 11 show the subjective listener performance on the intelligibility of the vocoded sentences. The third column is a repetition of the last column in Fig.10, and it represents the average word intelligibility on each of the sentences, while the fourth columns gives the average sentence intelligibility. A sentence was considered to be correct if all the words in that sentence were correctly transcribed by the listener. Therefore, the numbers in the fourth column must be less than or equal to those in the third column. While, from Fig. 11, one sees that listener performance was highly correlated with phoneme correctness, there were exceptions to the trend. For example, sentence number 5, where only 77% of the phonemes were transmitted correctly, was understood perfectly by all of the subjects. This is because the sentence was short and the phonetic errors were not on adjacent segments. Sentence 4, at 67% phoneme recognition, gives an indication of what would be unacceptable performance for a vocoder in a communications environment. It is clear from this and other experience we have had that phoneme accuracy should be no less than 70%.

There also seems to be another important variable controlling intelligibility. The subjects can be broken down

into two groups. Those who have had some experience listening to speech generated by the Synthesis-by-Rule (SBR) program, and those who have not. Fig. 12 indicates percent correct word recognition and sentence recognition in the first and second trials for subjects without experience and subjects with experience. The averages across all subjects who had had some experience with SBR speech were better on the first trial. After the first trial, the two groups seemed to perform about equally. This argues strongly that with a small amount of experience in listening to this type of vocoded speech, listeners will be much better able to understand the speech it produces. This means that the intelligibility results which should be considered to represent the projected results are those of the experienced listeners.

In conclusion, while marginal performance can be achieved with 70% phoneme recognition, we feel that at least 80% and preferably 85% phoneme recognition is needed for a communications task.

5. CONCLUSIONS FROM PILOT PROJECT

In this section we restrict our attention to experience with the 75 bps phonetic vocoder.

Because neither the acoustic-phonetic recognition program nor the synthesis-by-rule program were written specifically for the phonetic vocoder application, it was inevitable that compatibility and performance problems would arise. Below we summarize the main existing problems and then give conclusions as to what is needed to improve performance.

5.1 Existing Problems

The problems can be divided between the recognition and the synthesis components.

Recognition

(1) The recognition program was designed to produce a segment lattice with a number of likely phonetic choices for each segment. On the other hand, the phonetic vocoder required a single string of segments with a single best phonetic choice for each segment. This incompatibility was circumvented using a short-term solution to the problem.

(2) Though phonetic context was used at times in the recognition process, it was not used in a maximal manner. The result was that phonetic error degradation was not always

graceful, i.e., the recognized phoneme could be in error by more than one feature. On the other hand, the phonetic vocoder application requires graceful degradation for maximum intelligibility.

(3) The recognition program was designed to be fairly speaker independent. This was good for developing robust recognition, but it also resulted in a recognition accuracy that was less than it could have been. Recognition accuracy is of paramount importance in a phonetic vocoder.

(4) It was hard to tune the system to a new speaker, since it required the use of much data. For vocoder applications, it is desirable that the system be tuned with as little data as possible.

Synthesis

(1) The synthesis program was incompatible with the recognition program in more than one way. The synthesis program dealt with 50 distinct phonetic elements as compared to 71 in the recognition program. Also, the synthesis program was designed to generate its own pitch and duration information; it had to be modified to accept this information from the recognition program in order to maintain the natural intonation in the speech.

(2) The synthesis-by-rule program produced speech with good intelligibility but with a quality that was unnatural. The program required minimal storage but used a large set of rules. The lack of naturalness was due to the rule aspect of the program; there simply were not enough rules to capture the naturalness of the speech. Speech naturalness can be important for a phonetic vocoder application.

5.2 Conclusions

Our conclusions in terms of how to improve the performance of a future phonetic vocoder are as follows:

(1) Quantization and Encoding

The quantization and encoding of pitch and duration at a total rate of 2.75 bits/phoneme was judged to be adequate in maintaining the natural intonation of the speech. No major improvements are necessary.

(2) Recognition

(a) A best first choice recognition strategy is needed.

(b) The recognition should utilize more context information.

(c) Greater speaker dependence should be built in to achieve higher recognition accuracy.

(3) Synthesis

(a) Synthesis should be designed to be compatible with recognition.

(b) The resulting speech should be more natural.

(c) The program should include the possibility of synthesizing the voices of different speakers, by simply storing the right information.

6. RECOMMENDATIONS

Given our conclusions, as outlined in the previous section, we set out in the final part of the project to study, in a fair amount of detail, the feasibility of actually building a real-time, natural sounding 75 bps phonetic vocoder. A number of our staff with various types of expertise participated in this process in order to form realistic estimates and conclusions. The results of our deliberations are being written in a document that will contain a proposed system that would actually accomplish the desired goal. That document will be submitted to the sponsor at a later date.

The proposed phonetic vocoder aims to have the following features:

(1) 75 bps average transmission rate,

(2) 85-90% phonetic recognition,

(3) natural sounding synthesis,

(4) real-time operation using off-the-shelf technology.

The 75 bps data rate was already demonstrated in the pilot project. The 2.75 bits/phoneme allotted to pitch and gain were judged to be adequate in maintaining the natural intonation of the speech.

Although 80% phonetic recognition might be adequate for good intelligibility in context, we felt it was best to aim at 85%-90% recognition to ensure high intelligibility.

In listening to the output of the phonetic vocoder in the pilot experiment, the most striking characteristic of the synthesized speech was its lack of naturalness, its machine-like quality, which has been referred to sometimes as a "Donald Duck" quality. In spite of the fact that the speech was very intelligible, the "Donald Duck" aspect seems to dominate the general effect, especially for a new listener. Such degradation in quality is not a necessary feature of a phonetic vocoder. Natural sounding speech is possible. In our proposed system, we have outlined a synthesizer that would store more of the speaker's characteristics and use fewer rules to synthesize speech. The result is synthesized speech that sounds natural, without the "Donald Duck" effect.

Finally, with the use of off-the-shelf fast digital signal processing elements, it is possible to have the whole system run in real-time. This conclusion was based on actual counting of needed machine instructions and on the characteristics of existing signal processors. A complete system can be built in a period of 3 years, though an initial non-real-time system should be available earlier. It is foreseen that an effort of about 2.5 man-years per year would be needed to accomplish the task.

An effort to build such a system would serve as the culmination of several years of technology development by ARPA under the speech understanding and speech compression projects

separately, into a single real-time system. On-going hardware development in the speech compression project would make the real-world construction of a 75 bps vocoder an inexpensive reality in only a few years.

7. REFERENCES

[1] Berger, B., Rate Distortion Theory, Englewood Cliffs, NJ: Prentice-Hall, 1971.

[2] Makhoul, J., R. Viswanathan, and W. Russell, "A Framework for the Objective Evaluation Vocoder Speech Quality," IEEE-ICASSP, Philadelphia, Pa., pp. 103-106, April 1976.

[3] Viswanathan, R., J. Makhoul, and W. Russell, "Towards Perceptually Consistent Measures of Spectral Distance," IEEE-ICASSP, Philadelphia, Pa., pp. 485-488, April 1976.

[4] Kang, G.S. and D.C. Coulter, "600 BPS Voice Digitizer," IEEE-ICASSP, Philadelphia, Pa., pp. 91-94, April 1976.

[5] Woods, W.A. et al, "Speech Understanding Systems", Final Report, November 1974-October 1976, Volume II, BBN Report No. 3438, Bolt Beranek and Newman Inc., 1976.

[6] Makhoul, J., R. Viswanathan, L. Cosell, and W. Russell, "Natural Comunication with Computers: Speech Compression Research at BBN," Final Report, Vol. II, BBN Report No. 2976, Bolt Beranek and Newman Inc., Cambridge, Ma, Dec. 1974.

[7] Viswanathan, R. and J. Makhoul, "Specifications for ARPA-LPC System II," Network Speech Compression Note #82, Bolt Beranek and Newman Inc., Cambridge, Ma, Feb. 1976.

[8] Viswanathan, R., "Variable Frame Rate Transmission of Pitch and Gain," Network Speech Compression Note #96, Bolt Beranek and Newman Inc., Cambridge, Ma, Sept. 1976.

[9] Blackman, E., R. Viswanathan, and J. Makhoul, "Variable-to-Fixed Rate Conversion of Narrowband LPC Speech," to be presented at the 1977 ICASSP, Hartford, Ct, May 9-11, 1977.

[10] Itakura, F., "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. ASSP, pp. 67-72, Feb. 1975.

[11] Huffman, D.A., "A Method for the Construction of Minimum-Redundancy Codes," Proc. IRE, Vol. 40, pp. 1098-1101, Sept. 1952.

[12] Makhoul, J., "Linear Prediction: A Tutorial Review," Proc. of the IEEE, Vol. 63, No. 4, April 1975.

[13] Viswanathan, R., J. Makhoul, and R. Wicke, "The Application of a Functional Perceptual Model of Speech to Variable-Rate LPC Systems," to be presented at the 1977 ICASSP, Hartford, Ct, May 9-11, 1977.

[14] Schwartz, R.M. and V.W. Zue, "Acoustic-Phonetic Recognition in BBN SPEECHLIS," presented at the 1976 ICASSP, Philadelphia, Pa, April 1976.

[15] Cook, C. and R. Schwartz, "Advanced Acoustic Techniques in in Automatic Speech Understanding," to be presented at the 1977 ICASSP, Hartford, Ct, May 9-11, 1977.

[16] Woods, W.A. et al, "Speech Understanding Systems," Final Report, November 1974-October 1976, BBN Report No. 3438, Vols I-V, Bolt Beranek and Newman Inc., Cambridge, Ma, 1976.

[17] Klatt, D.H., "Structure of the Phonological Component for a Synthesis-by-Rule Program," IEEE Trans. ASSP-24, pp. 391-398, 1976.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: BBN REPORT NO. 3408

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): A FEASIBILITY STUDY OF VERY LOW RATE SPEECH COMPRESSION SYSTEMS.

5. TYPE OF REPORT & PERIOD COVERED: Final Report. 19 July 1976 - 18 Jan 1977

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): J. Makhoul, C. Cook, R. Schwartz, D. Klatt

8. CONTRACT OR GRANT NUMBER(s): MDA 903-75-C-0180, ARPA Order-3253

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Bolt Beranek and Newman Inc. 50 Moulton St, Cambridge, MA 02139

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS

12. REPORT DATE: February 1977

13. NUMBER OF PAGES: 55

14. MONITORING AGENCY NAME & ADDRESS (If different from Controlling Office): 59p.

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce for sale to the general public.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, If different from Report)

18. SUPPLEMENTARY NOTES

This research was supported by the Defense Advanced Research Projects Agency under ARPA Order No. 3253.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

speech analysis, speech recognition, speech synthesis, speech synthesis-by-rule, speech compression, vocoders, very low rate vocoders, formant vocoders, LPC vocoders, variable-rate vocoders

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report documents the results of a feasibility study for very low rate speech compression systems. Two systems were implemented: (1) A variable rate, hybrid formant-LPC vocoder transmitting at an average of 500 bps, and (2) a phonetic vocoder transmitting at an average rate of 75 bps. The intelligibility of the 500 bps system was judged to be high, but with a loss in naturalness and speaker identifiability. The 75 bps system utilized a phonetic recognition program and transmitted phonemes, pitch and duration. The synthesis at the → next page

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

060100

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

LB

receiver was performed using a synthesis-by-rule program, with a resulting good intelligibility but machine-like quality. In a listener judgment experiment of the vocoder, a recognition accuracy of over 80% was deemed necessary for communication purposes. A study was then undertaken to determine the feasibility of a real-time 75 bps vocoder. It was concluded that such a system was indeed feasible, with over 80% phonetic recognition, natural sounding synthesis, and real-time operation using off-the-shelf technology.