1976-10-01

# Two preliminary studies of the intelligibility of predictor-coefficient and formant-coded speech

L. Keeler
lokeeler@cox.net

G. Clement

W. Strong

E. Palmer

Note that

$$\lim_{M \to \infty} |T - T_c| = 0$$

implying that the DCT is asymptotically equivalent [4] to the Karhunen-Lóeve transform (KLT) of Markov-1 processes. Moreover, since for large $M$ and $\rho \neq 1$, we have

$$|T - T_c| = \sqrt{2}\,\frac{\rho}{1-\rho^2}\,0(M^{-1/2})$$

we conclude that the degradation in performance in filtering and coding [3] vanishes like $M^{-1/2}$.

An asymptotic equivalence between the KLT and the DCT was also argued by Shanmugam [5] using a circulant extension of $T$. His argument, however, remains incomplete as the relation between the DCT and the circulant matrices used in [5] is rather unclear.

In order to calculate $|T - T_F|^2$, recall [3] that for $T_{ij} = t(|i-j|)$ we have

$$(T - T_F)_{ij} = \frac{|i-j|}{M}\,[t(|i-j|) - t(M - |i-j|)]$$

and substituting $t(|i-j|) = \rho^{|i-j|}$ we obtain

$$M|T - T_F|^2 = \frac{2\rho^2(1+\rho^{2M})}{(1-\rho^2)^2} - \frac{2(1+\rho^2)\rho^2(1-\rho^{2M})}{M(1-\rho^2)^3} - \frac{\rho^M(M^2-1)}{3}.$$

It shows that the asymptotic behavior of $|T - T_F|$ for large $M$ is identical to that of $|T - T_c|$. Thus, the performance difference between the DCT and the DFT must vanish like $M^{-1}$. Indeed, for large $M$ one obtains the positive difference

$$|T - T_F|^2 - |T - T_c|^2 \cong \frac{4\rho^2}{M^2(1-\rho^2)(1+\rho)^2} \qquad \rho < 1,$$

indicating that the cosine transform is closer to optimal than the Fourier transform over the entire range of $0 < \rho < 1$.

For moderate values of $M$ we should examine the expressions for $|T - T_F|^2$ and $|T - T_C|^2$ over the range $0 \leqslant \rho \leqslant 1$. The two are plotted, in a normalized form, in Fig. 1. We chose $|T - I|^2$ as a common normalizing factor, where $I$ is the identity matrix, and so

$$|T - I|^2 = \frac{2\rho^2}{M(1-\rho^2)^2}\,[M - 1 - M\rho^2 + \rho^{2M}].$$

It measures the degree of cross correlation contained in the unprocessed signal, and, therefore, the maximum amount of decorrelation that can be accomplished by any transform (i.e., the KLT). The ratio

$$\frac{|T - T_U|^2}{|T - I|^2}$$

represents the fractional correlation left "undone" by a transformation $U$.

Fig. 1 shows that for $M = 8$, 16, 64, and for the entire range of $0 < \rho < 1$, $|T - T_F|^2$ is higher than $|T - T_C|^2$. The difference between the two are quite noticeable, occasionally reading a ratio of 2:1.

## CONCLUSIONS

We established that the DCT is asymptotically equivalent to the KLT of Markov-1 signals and demonstrated that the rate of convergence is on the order of $M^{-1/2}$. $|T - T_C|^2$ is shown to be smaller than $|T - T_F|^2$ for all values of $M$ and $\rho$, i.e., the discrete cosine transform offers a better approximation to the KLT of Markov-1 signals than the DFT.

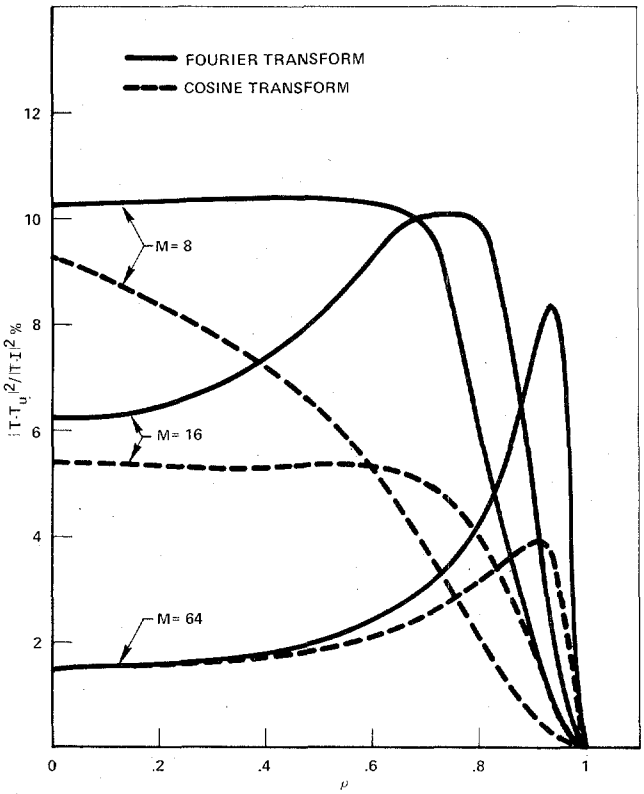


Fig. 1. Normalized correlation measures for Fourier (solid lines) and cosine (broken lines) transforms.

## REFERENCES

[1] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.* (Corresp.), vol. C-23, pp. 90–93, Jan. 1974.

[2] R. W. Means, H. J. Whitehouse, and J. M. Speiser, "Television encoding using a hybrid discrete cosine transform and a differential pulse code modulator in real time," in *Proc. IEEE Nat. Telecommunications Conf.*, San Diego, CA, Dec. 1974.

[3] J. Pearl, "On coding and filtering stationary signals by discrete Fourier transforms," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-19, pp. 229–232, Mar. 1973.

[4] —, "Asymptotic equivalence of spectral representations," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 547–551, Dec. 1975.

[5] K. S. Shanmugam, "Comments on discrete cosine transform," *IEEE Trans. Comput.* (Corresp.), vol. C-24, p. 759, July 1975.

# Two Preliminary Studies of the Intelligibility of Predictor-Coefficient and Formant-Coded Speech

LYNN O. KEELER, GARY L. CLEMENT, WILLIAM J. STRONG, AND E. PAUL PALMER

*Abstract*—**Two preliminary studies comparing the intelligibilities of predictor-coefficient versus formant-frequency-coded speech and the intelligibilities of predictor-coefficient-coded speech using different numbers of coefficients are reported.**

## INTRODUCTION

Much of the motivation for research in speech analysis–synthesis stems from a desire to better understand the essential

Manuscript received July 16, 1975; revised January 9, 1976.

The authors are with the Department of Physics and Astronomy, Brigham Young University, Provo, UT 84602.

properties of speech. These properties may then be exploited for such things as efficient transmission and storage of speech signals, computer speech recognition, speaker verification and identification, machine synthesis of speech, and aids for the deaf. Information rate is of primary concern when considering speech analysis–synthesis systems for the efficient transmission and storage of speech. However, in other applications such as machine recognition or machine synthesis of speech, other measures of economy such as the total number of parameters needed to specify the signal may be of more interest than the information rate *per se*. These preliminary studies are directed to the latter consideration.

Recent speech analysis–synthesis research has included the use of linear prediction methods [1], [5]–[7] and formant methods [8], [10]. The choice of which parameter set to use in coding speech may depend upon the use to be made of the speech code since a code useful in one application may be less useful in another. Often, a choice between possible parameter sets can be based on the criteria of efficiency of computation, efficiency of the representation, and completeness of the representation.

If the final use of the speech code is to produce speech, an obvious test of its adequacy is to test the intelligibility of the speech with human listeners. For a code which may not be used to produce speech, but is used in something such as machine recognition, the value of intelligibility testing is less obvious. However, short of testing the code in its final application, intelligibility testing seems to be a useful method for gaining information about the adequacy of a code. Several different intelligibility testing methods have been developed some of which use closed response sets of rhyming words [4], [11]. These tests have the advantage of giving scores which are related to the ability of a system to perform in communication use and of providing diagnostic information about the confusions found in the system. Also, they require little test-crew training.

This correspondence describes two preliminary studies comparing the intelligibility of predictor-coefficient and formant-coded speech. The two studies are somewhat disparate in nature because they used different rhyming word tests due to their having been conducted at different times. The first study describes a comparison of natural speech and speech synthesized using twelve or six predictor coefficients, and five or three formant frequencies and amplitudes. A modified rhyme test (MRT) of 250 words employing rhyming minimal contrasts [4] was used to assess the intelligibilities of the natural speech and the four kinds of synthetic speech. The second study describes a comparison of speech synthesized using two, four, six, eight, ten, or twelve predictor coefficients. Form IV of the diagnostic rhyme test (DRT) employing 96 rhyming word pairs [11] was used to assess the following consonantal attributes: voicing, nasality, sustention, sibilation, graveness, and compactness. Only one male talker was used in each study (a different one in each), and hence the results to be reported must be taken as tentative.

## Description of Experiments

Speech was recorded in an anechoic chamber with the talker attempting to speak with constant vocal effort. After a short practice period, the words of the articulation test were recorded, then low-pass filtered at 4.5 kHz, sampled at 10 kHz, and stored on magnetic disk for processing. An autocorrelation, linear-prediction, analysis–synthesis method similar to that of Markel [7] was used. The autocorrelation coefficients and then the predictor coefficients were computed for data points windowed with a 256-point Hamming window. The unwindowed data were used for pitch-period calculations [3] and to obtain zero crossings and slope changes for use in the voicing decision. A gain factor was also computed for use in synthesis. A new analysis was performed at 10-ms increments in the speech signal, and the suitably normalized parameters were stored on disk for use in synthesis. In synthesis, for a voiced decision, a pulse excitation was used; for a noise decision, noise excitation was used; and for a mixture decision, a combination of pulse and noise excitation was used. For the voiced and mixture decision, the predictor coefficients and other parameters were changed pitch synchronously, while for the noise decision they were changed every 10 ms in the first study and every millisecond in the second study.

The formant-frequency analysis method used was basically that of Christensen *et al.* [2]. Predictor coefficients were used to calculate a spectrum of the speech sample analyzed, and the second difference of the smoothed spectrum was calculated. The peaks in the negative of the second difference were found and their frequency positions stored. The largest amplitudes in the spectrum which corresponded to these frequencies were used to determine the formant frequencies which, along with formant amplitudes, were stored on disk. The synthesis program employed a parallel-pole synthesizer [10]. For a voiced decision, a pulse excitation was used; for a noise decision, the excitation was noise; and for a mixture specification, the first filter was pulse excited while the other filters were noise excited. The outputs of adjacent filters were added out of phase.

In study one, a reference tape of the 250-word modified rhyme test was made up of normal speech that had been processed through tape recording and analog-to-digital and digital-to-analog conversion. In addition, synthetic speech was generated using twelve and six predictor coefficients (12 PC and 6 PC) and five and three formant frequencies and amplitudes (5 FFA and 3 FFA). Ten different listeners were used with each version of the speech so that each listener heard each of the 250 words only once, and learning effects were thus eliminated. In the test, the word thought to be heard was chosen from a set of five rhyming words printed on the test sheet.

In study two, comparing speech synthesized using twelve, ten, eight, six, four, and two predictor coefficients, the 192 words (96 rhyming pairs) of the diagnostic rhyme test were presented and the person made a choice of which word he heard from each pair listed on the test sheet. Ten listeners took the test in which each word was presented twice in each version of the speech. Word order and speech version were randomized.

During the tests, the subjects were seated in an anechoic room at a radius of about 15 ft in front of an Electro-Voice Sentry III loudspeaker such that no listener was more than 30° off-axis. The average loudness level of the speech was estimated as 75 DBA at the listener's position using a sound level meter. The signal-to-noise ratio of the speech was estimated to be approximately 30 dB. No explicit attempt was made to balance the juries with respect to age, sex, etc., although all subjects were college undergraduates or graduate students with no known hearing defects.

## Results and Discussion

*Study One–Predictor Coefficients versus Formants:* The intelligibility scores for the original speech and the four types of synthetic speech are given in Table I. The "all consonants" column shows the percentage of all consonants (initial and final) received correctly out of 2500 possible responses (250 words times 10 listeners). The "initial consonants" column shows the percentage of initial consonants received correctly out of 1250 possible responses and the "final consonants" column shows the percentage of final consonants received correctly out of 1250 possible responses. Test 6 in the table is from a similar study in which four formant frequencies and amplitudes were used [10].

TABLE I
INTELLIGIBILITY SCORES FOR SIX TESTS

| Test | Condition | Per cent correct responses: All consonants | Initial consonants | Final consonants | Standard deviation |
|---|---|---|---|---|---|
| 1 | Normal | 96 | 96 | 97 | 1.3 |
| 2 | 12 Predictor coefficients | 88 | 87 | 89 | 2.5 |
| 3 | 6 Predictor coefficients | 77 | 78 | 77 | 3.9 |
| 4 | 5 Formants | 86 | 84 | 89 | 1.9 |
| 5 | 3 Formants | 85 | 85 | 86 | 1.7 |
| 6 | 4 Formants (Strong 1967) | 84 | 82 | 86 | --- |

Although the 12-predictor-coefficient speech was of apparent high quality and very natural sounding from a subjective point of view, it showed a decrease in intelligibility of 8 percent from the normal speech (which is statistically significant). The intelligibility scores for speech synthesized using 12 predictor coefficients, five formant frequencies and amplitudes, and three formant frequencies and amplitudes, were nearly the same. This seems to indicate that in coding for the purpose of synthesizing speech, the three methods are nearly equivalent and little additional information for human speech recognition is gained by specifying five formant frequencies and amplitudes instead of three in the region of 0–5 kHz. The results here are in good agreement with the earlier study by Strong.

The data for initial consonants are presented in a different manner in Fig. 1 for comparison with study two later. Errors in the speech attributes of voicing, nasality, sustention, sibilation, graveness, and compactness are shown for each of the speech versions. The Griffiths' rhyme test used in this study does not lend itself readily to the separate determination of errors for the six attributes listed because confusions among the rhyming words can involve more than one attribute. When the errors involved more than one attribute, errors were arbitrarily divided equally among each attribute involved. This has resulted in an inflation of errors for the voicing attribute where errors are rare and a corresponding deflation of errors for other attributes.

The errors in sustention may be related to the rate at which the speech parameters are changed, 10 ms in this study. Errors in graveness and compactness may be attributable to the fact that the approximations to the speech spectrum made by the systems used were not close enough to that of real speech; an increase in the number of errors in graveness and compactness is seen in the 6 PC speech as compared to 12 PC speech. Also, since formant transitions are known to occur rapidly in some instances [7], a 10-ms parameter changing rate may not be sufficient to track them.

*Study Two—Various Numbers of Predictor Coefficients:* The percentage intelligibility scores for the original speech and the six types of predictor-coefficient coded speech are given in Table II along with a breakdown of the overall scores for the six consonant attributes of voicing, nasality, sustention, sibilation, graveness, and compactness. The standard deviation is given in parentheses following each score. The results for this study have been corrected for guessing by subtracting the number wrong from the number right. The intelligibility scores for each consonant attribute and for the total are plotted versus the version of speech in Fig. 2.

Results for the original speech indicate high intelligibility with the exception of the attribute graveness. Voiers *et al.* [11] indicate that the apprehensibility of graveness and sustention tends rather generally to be the most difficult and most susceptible to speech degradation, while the apprehensibility of voicing and nasality tends to remain higher under most conditions. Subjectively, the 12-PC speech was highly intelligible and not significantly different from real speech. All the other versions of the synthetic speech were significantly different ($p < 0.001$) from the real speech in overall scores. The groupings of the scores are interesting with 12-PC and 10-PC speech together, 8-PC, 6-PC, and 4-PC speech grouped, and 2-PC speech in a group by itself.

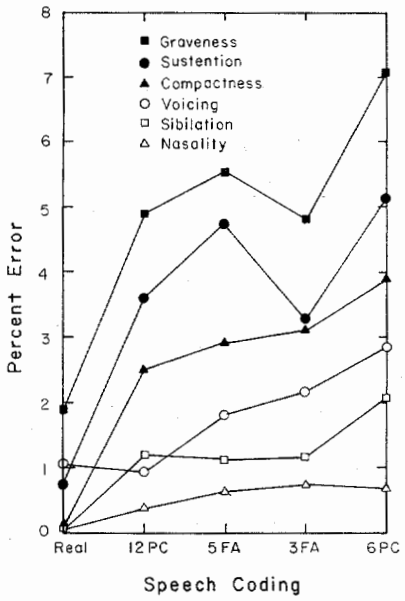


Fig. 1. Errors in the apprehension of initial consonants due to the attributes of voicing, nasality, sustention, sibilation, graveness, and compactness for different speech codings.

Voicing and nasality remain high over all speech versions. It is somewhat surprising that they are not degraded more for low numbers of predictor coefficients since spectra of unvoiced and nasal sounds usually include zeros as well as poles [1], while linear prediction uses an all-pole approximation to the original spectrum in all cases.

The difference between real and 12-PC speech is most apparent for sustention, and this may be related to the frame rate used in the analysis–synthesis process. Acoustically, the distinction between a sustained versus an interrupted sound is essentially whether the onset of energy is gradual or abrupt. Abrupt changes in energy will tend to be smoothed out by the windowing process. This may be the reason for the initial decrease in the apprehensibility of sustention since further decreases are small except for 2-PC speech where the decrease may be due to severe degradation of the spectrum.

Graveness and compactness are related to place of articulation and are acoustically correlated with the relative positions of the second and third formants. The apprehensibility of these attributes would thus be expected to depend quite strongly on having a well-defined spectrum. It is not too surprising that these two attributes fall off more rapidly than any of the others as the spectral approximation is degraded. Since 2-PC speech has no more than one formant, most of the information about the relative positions of the second and third formants in the original speech is lost which may account for the very large decrease in apprehensibility of these attributes for 2-PC speech.

## SUMMARY

The 12-predictor-coefficient, five-formant, and three-formant codings produce comparable overall intelligibilities that are

TABLE II
INTELLIGIBILITY SCORES FOR SEVEN TYPES OF SPEECH

| | Percent Correct Responses | | | | | | |
|---|---|---|---|---|---|---|---|
| Condition | Voicing | Nasality | Sustension | Sibilation | Graveness | Compactness | Overall |
| 2-PC | 95.6 (6.3) | 94.4 (6.6) | 67.8 (11.5) | 88.8 (9.5) | 21.6 (13.5) | 51.6 (8.5) | 69.9 (6.3) |
| 4-PC | 98.1 (2.2) | 96.9 (5.1) | 81.2 (12.2) | 95.6 (6.6) | 63.1 (10.1) | 82.5 (7.5) | 86.2 (4.8) |
| 6-PC | 98.1 (3.0) | 96.9 (3.6) | 81.6 (13.0) | 96.9 (3.3) | 68.1 ( 9.4) | 83.4 (8.3) | 87.5 (3.9) |
| 8-PC | 98.1 (3.0) | 96.9 (3.9) | 80.3 (10.8) | 95.9 (3.9) | 68.1 ( 8.4) | 83.4 (6.3) | 87.1 (4.4) |
| 10-PC | 98.8 (2.2) | 97.5 (5.1) | 85.3 ( 7.4) | 96.9 (2.6) | 75.9 ( 8.6) | 91.2 (3.2) | 90.9 (2.3) |
| 12-PC | 98.4 (2.2) | 99.1 (1.5) | 85.9 (8.0) | 98.1 (3.0) | 79.1 ( 7.1) | 91.2 (3.6) | 92.0 (3.3) |
| Real | 99.1 (1.1) | 98.9 (1.8) | 95.8 (3.5) | 98.0 (2.5) | 85.8 ( 5.4) | 94.7 (2.7) | 95.3 (1.8) |

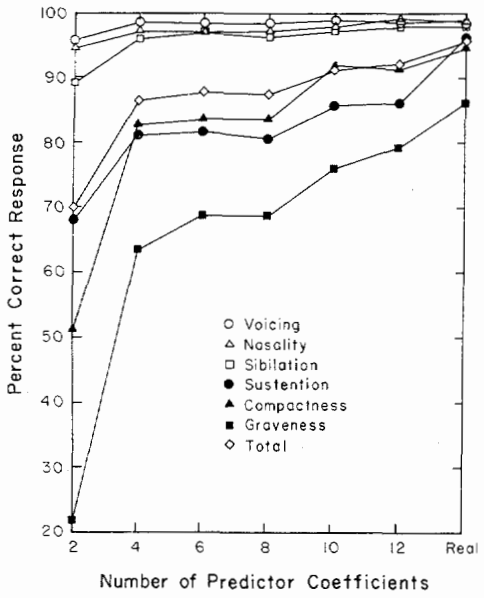


Fig. 2. Percentage correct response versus number of predictor coefficients for various speech attributes.

better than those of six-predictor-coefficient coding. The three-formant coding is a more economical representation than either the 12-predictor-coefficient or five-formant codings. It is as economical as six-predictor-coefficient coding, but is more complete (85 percent versus 77 percent intelligibility). (These economies are considered in terms of total number of parameters and do not supply information on efficiencies in terms of bit rates for communication systems.) The formant codings are less efficient to compute than the predictor-coefficient codings.

Increasing the number of predictor coefficients increases the intelligibility and decreases the economy of the representations in an essentially monotonic fashion. The major break in the intelligibility curve occurs between two and four predictor coefficients.

Both studies show that the attributes of voicing, nasality, and sibilation are rather robust, whereas sustention, graveness and compactness are more susceptible to degradation. The results of the two studies cannot be compared directly because different talkers and different tests were used in the two cases. However, the higher overall scores of the DRT compared with those of the MRT are consistent with results reported by Smith [9]. The results are consistent internally and with those of other studies, although they must be taken as tentative since only one male talker was used in each study.

## REFERENCES

[1] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, 1971.
[2] R. L. Christensen, W. J. Strong, and E. P. Palmer, "A comparison of three methods of extracting resonance information from predictor-coefficient coded speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 8–14, Feb. 1976.
[3] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Amer.*, vol. 46, pp. 422–448, 1969.
[4] J. D. Griffiths, "Rhyming minimal contrasts: A simplified diagnostic articulation test," *J. Acoust. Soc. Amer.*, vol. 42, pp. 236–241, 1967.
[5] F. Itakura and S. Saito, "Analysis-synthesis telophony based on the maximum likelihood method," presented at the 6th Int. Congr. Acoustics, Tokyo, Japan, Paper C-5-5, 1968.
[6] J. Makhoul and J. J. Wolf, "Linear prediction and the spectral analysis of speech," Bolt Beranek and Newman, Inc., Cambridge, MA, Rep. 2304, 1972.
[7] J. D. Markel, "Digital inverse filtering–A new tool for formant trajectory estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 129–137, June 1972.
[8] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, pp. 634–648, 1970.
[9] S. P. Smith, "Comparison of two diagnostic intelligibility test methods," *J. Acoust. Soc. Amer.*, vol. 56, p. S52, 1974.
[10] W. J. Strong, "Machine-aided formant determination for speech synthesis," *J. Acoust. Soc. Amer.*, vol. 41, pp. 1434–1442, 1967.
[11] W. D. Voiers, A. D. Sharpley, and C. J. Hehmsoth, "Research on diagnostic evaluation of speech intelligibility," Tracor, Inc., Austin, TX, 1973.

## Comments on "A Fast Algorithm for the Estimation of Autocorrelation Functions"

HUGH LARSEN

*Abstract*–**A recently published algorithm for the estimation of arithmetic autocorrelation functions may be further refined. The pre-multiplications may be converted into simple additions which are**

Manuscript received December 4, 1975; revised April 20, 1976. This research was supported in part by NASA under Grant 46-001-041-1.

The author is with the Department of Electrical Engineering, University of Vermont, Burlington, VT 05401.