# CODING OF THE SPEECH SPECTRUM IN THREE TIME-VARYING SINUSOIDS [a]

Robert E. Remez

*Department of Psychology*
*Barnard College*
*Columbia University*
*New York, New York 10027*

Philip E. Rubin

*Haskins Laboratories*
*New Haven, Connecticut 06510*

David B. Pisoni

*Department of Psychology*
*Indiana University*
*Bloomington, Indiana 47405*

## INTRODUCTION

A range of alternatives now exists for coding the rich spectrum of speech signals. Some techniques create highly intelligible speech by producing signals from digital records of sampled natural utterances. Other techniques, like those used in formant synthesizers and LPC-vocoders, may generate intelligible though less natural-sounding speech from schematic representations of the essential correspondences between acoustic structure and phonetic properties. Perceptual experiments have found that some acoustic impostors of speech are intelligible, while others are not. Intelligible synthesis has been said to retain those acoustic properties of natural speech that are necessary for perception, and those that are unintelligible are said to discard at least some of them. This characterization of intelligibility stated in perceptual terms seems to mean that the identification of phonetic sequences is based just on those acoustic elements, or cues, that bear phonetic information. In practice, however, the cues prove elusive to define. The phonetic value of any particular extract of the natural signal may vary widely, depending on contextual factors. Except by hindsight, it becomes difficult to state precisely which elements in a given signal possess phonetic value and which do not. Overall, then, it seems self-evident that the intelligible simplifications of the signal incorporate the acoustic properties of natural speech upon which perception relies. But the definition of those essential acoustic properties remains a wholly empirical matter.

A number of classic experiments [1] show that successful recodings of the speech signal need not contain every acoustic detail of natural utterances. Nevertheless, our practice of describing the apparently relevant acoustic details as particular short-time spectra or as particular transitions of formants in

[a] This work was supported by Grant HD 15672 (to R.E.R.) and Grant HD 01944 (to Haskins Laboratories) from the National Institute of Child Health and Human Development, and by Grant MH 24027 (to D.B.P.) from the National Institute of Mental Health.



0077-8923/83/0405-0485 $1.75/0 

regions of the spectrum has led us to overlook the potential contribution of phonetic information from *coherent* spectrum *variation*. Our recent experiments with sinusoidal replicas of speech signals address this topic, and the results of these tests indicate that coherent variation of acoustic patterns can provide sufficient phonetic information for the linguistically competent listener. Moreover, these studies revealed that speech perception can be reliable given a stimulus possessing none of the first-order properties of natural speech. The implications for the development of coding schemes to use in intelligent prostheses are quite explicit, or so it would seem: the time-varying coherence of the spectrum bears phonetic information, apparently independent of the elements that compose the dynamic patterns.

## Perception of Sinusoidal Signals

In our studies [2-5] we have used signals consisting of three time-varying sinusoids, each of which varies in a formant-like manner. We fabricate each sinusoidal pattern by computing the resonant center-frequencies of a natural utterance, using the analysis technique of linear prediction.[6] The table of values produced through this analysis is used to set frequency and amplitude parameters of a sinewave synthesizer. Typically, three tones are synthesized, each imitating the frequency and amplitude changes of one of the first three formants. Sinusoidal tone-complexes lack fundamental frequency, harmonic spectrum, and broadband formants (the short-time characteristics of natural speech), although there *is* energy, albeit infinitely narrowband, at the computed resonance peaks throughout the duration of each pattern. In consequence, the time-varying properties of the sinewave pattern, specifically the coherence of the *changes* of the energy peaks over time, replicate the natural case.

The perceptual effects of sinewave stimuli were easy to predict. Because the short-time spectra of three-tone signals differ drastically from natural and even synthetic speech; because no talker is capable of producing three simultaneous "whistles" with these bandwidths, in this frequency range;[7] and because the frequency and amplitude changes of the tones are not synchronized, the perceiver should hear three independent streams, one for each sinusoid. The perceiver should hear no phonetic qualities.

However obvious this prediction seemed, there was an equally plausible, though contrasting, prediction. Suppose that the listener was able to disregard the short-time differences between sinusoidal signals and speech, and could attend, instead, to the overall pattern of change of the three tones. The pattern of change of the frequency peaks resembles the resonance changes produced by the vocal tract when articulating speech. If the listener can apprehend this coherence in the time-varying properties of the nonspeech signal, then he should hear a phonetic message spoken by an impossible voice.

Given nonspeech stimuli whose time-varying properties are analogous to vocal signals, listeners perceived the signals in *both* of the ways we predicted. Those listeners who were told nothing about the stimuli heard science-fiction-like sounds, electronic music, sirens, computer bleeps, and radio interference. Those listeners who instead were instructed to transcribe a "strangely synthesized English sentence" did exactly that—they identified the unnatural "voice" quality of the patterns, but transcribed the patterns as they would have the original utterances upon which we based the sinewave stimuli.[2]



In addition to this manipulation of instructions, we also varied the particular stimulus between the groups of listeners. Seven stimulus conditions were used in all, in which subjects heard either three tones, or tones in pairs, or the individual tones. The results of this manipulation indicated that the subjects demanded very specific stimulus structure to detect the phonetic message. Only the three-tone combination and the pair of T1 and T2 (matched respectively to the first and second formants) were transcribed correctly. None of the single-tone patterns was judged to contain any of the words, even when listeners knew which words to expect. We might conclude from this that single formants do not possess sufficient information for phonetic perception.

These studies indicate that speech perception is possible despite drastic departures from the short-time spectra of natural speech—despite absence of broadband formants, harmonic spectrum, and fundamental frequency—insofar as the time-varying properties of speech signals are preserved; and, insofar as the listener is able to attend to the coherent time-variation of the acoustic pattern. Evidently, both of these general qualifications must be met for phonetic perception of sinusoids to occur, for listeners who were not warned to expect speech did not for the most part hear phonetic sequences in tones.

## Effects of Phone Class, Amplitude Variation, and Temporal Variation on Intelligibility

Although we have found that sinusoidal replicas of naturally produced utterances are less intelligible, overall, than is speech produced by a conventional terminal analog process, this differential effect does not appear to depend on the recalcitrance of particular phone classes. We might have initially suspected that voiceless stops and fricatives would be unproducible by this technique. On the contrary, we found that those kinds of segments are recognized without benefit of semantic or contextual constraints, as in the sentence "Will Doctor Bronstein meet Thornton in Winnipesaukee." Note that nasals, liquids, stops, voiced consonants and vowels are all susceptible to this kind of exclusively time-varying code.

It seems, too, that listeners do not glean much information from the amplitude variation of the tones, nor do they particularly mind grossly inappropriate amplitude values. If the sinusoidal sentences were characterized as "impoverished" stimuli, then we would have expected that inappropriate amplitude values would degrade perception. The absence of other acoustic elements resembling the natural-signal constituents should have forced listeners to rely on whatever other acoustic structure was available, and relative amplitude has been correlated not only with syllable structure, but with vowel and consonant identity as well. However, coherent frequency variation appeared to be sufficient. In fact, with sinusoidal signals the rate of variation is crucial for perception. In a study in which the rate of frequency variation of three tones was manipulated, sinewave sentences were not intelligible at rates of frequency change that departed more than a factor of two from the natural rate of variation. These effects indicate, first, that the normal listener may ordinarily make extensive use of time-varying information during ordinary speech perception. Second, the sufficiency of the complex three-tone carrier holds the prospect for a device that conveys the set of phonetic segments without the necessity of reproducing the entire rich spectrum of natural signals.

## Problems with the Technique

Despite the evidence we have discussed in favor of the view that phonetic information is adequately conveyed by the time-varing spectrum, we should point out several potential drawbacks with this method. Because no tone component follows the fundamental frequency contour (our tests show that a fourth tone tracking FØ does not fuse with the phonetic percept), this acoustic aspect is simply omitted from transmission. In consequence, listeners do not experience the sinusoidal voice pitch to be modulated in the same way a natural voice is, and our tests show that Tone 1 is doubling as acoustic information for manner and intonation. Prosodically, sinusoidal speech is unusual, and distractingly so to the listener.

Additionally, sinusoidal replicas of speech are not uniformly highly intelligible, as our tests have shown. Typically, we present a series of sentences, each one repeated three or four times. Listeners then transcribe what they hear. Perhaps because of the strangeness of the sinusoidal voice, at least 30% of our typical, normal adult subjects hear no phonetic segments. Eliminating them from the group improves intelligibility statistically, but only to about 65% correctly identified words. Clearly, we must determine the effect of familiarity with sinusoidal signals on transcription, if there is any effect to speak of. Research by Grunke and Pisoni [8] is encouraging on this score. In their study, naive listeners learned to classify sinusoidal "syllables" composed of tones. Subjects labeled the tone patterns either with acoustic symbols that referred to the tone properties of spectrum changes (for example, rising, falling, steady) or with phonetic labels appropriate for the phonetic value each pattern was thought to have (for example, ba, da). Acoustic labeling was more reliable for the single tones and for double-tone patterns analogous to the second and third formants. Essentially, when conditions for phonetic perception of tones were not satisfied, acoustic labels were easier to learn. However, three-tone patterns were more reliably classified with the phonetic rather than acoustic terms. This indicates that the stimulus properties themselves facilitate the subject's attention to phonetic information. The effect of familiarity with three-tone signals on phonetic perception of sinusoids deserves further attention.

To conclude, our studies of speech signals replicated with tonal patterns has revealed that time-varying structure alone can convey phonetic information. While we particularly hope that our pursuit of this issue will resolve the paradoxes of isolated-cue conceptualizations of the information in speech signals, it is not uninteresting to speculate whether this coded simplification of the rich spectrum of speech might not present a possible technique for use in cochlear implants. Even in view of the basic questions that remain for us, we contribute the results to our tests with the normal listener to this end.

## Summary

Recent perceptual experiments with normal adult listeners show that phonetic information can readily be conveyed by sinewave replicas of speech signals. These tonal patterns are made of three sinusoids set equal in frequency and amplitude to the respective peaks of the first three formants of natural-speech utterances. Unlike natural and most synthetic speech, the spectrum of sinusoidal patterns contains neither harmonics nor broadband formants, and is indentified as grossly unnatural in voice timbre. Despite this drastic recoding of the short-time speech spectrum, listeners perceive the phonetic content if the temporal properties of spectrum variation are preserved. These observations suggest that phonetic perception may depend on properties of coherent spectrum variation, a second-order property of the acoustic signal, rather than any particular set of acoustic elements present in speech signals.



## Acknowledgments

We wish to thank Louis Goldstein and Michael Studdert-Kennedy for their helpful, relentless criticism.

## References

1. Liberman, A. M. & M. Studdert-Kennedy. 1978. Phonetic perception. *In* Handbook of Sensory Physiology. Vol. VIII: Perception. R. Held, H. Leibowitz, and H.-L. Teuber, Eds.: 143–178. Springer Verlag. New York, NY.
2. Remez, R. E., P. E. Rubin, D. B. Pisoni & T. D. Carrell. 1981. Speech perception without traditional speech cues. Science **212:** 947–950.
3. Remez, R. E., P. E. Rubin & T. D. Carrell. 1981. Phonetic perception of sinusoidal signals: Effects of amplitude variation. J. Acoust. Soc. Am. **69:** S114.
4. Remez, R. E. & P. E. Rubin. Phonetic perception of sinusoidal signals: Effects of temporal variation. Paper presented at the Twenty-Second Annual Meeting of the Psychonomic Society, November 14, 1981.
5. Remez, R. E. & P. E. Rubin. 1982. Perception of voice pitch in sinusoidal imitations of speech. J. Acoust. Soc. Am. **71:** S96.
6. Markel, J. D. & A. H. Gray, Jr. 1976. Linear Prediction of Speech. Springer Verlag. New York, NY.
7. Busnel, R. G. & A. Classe. 1976. Whistled Languages. Springer Verlag. New York, NY.
8. Grunke, M. E. & D. B. Pisoni. 1982. Perceptual learning of mirror-image acoustic patterns. Percept. Psychophys., **31:** 210–218.