Estimating speech spectra for copy synthesis by linear prediction and by hand

Robert E. Remez,^{a)} Kathryn R. Dubowski, Morgana L. Davids, Emily F. Thomas, Nina U. Paddu, Yael S. Grossman, and Marina Moskalenko

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

(Received 22 July 2010; revised 21 July 2011; accepted 9 August 2011)

Linear prediction is a widely available technique for analyzing acoustic properties of speech, although this method is known to be error-prone. New tests assessed the adequacy of linear prediction estimates by using this method to derive synthesis parameters and testing the intelligibility of the synthetic speech that results. Matched sets of sine-wave sentences were created, one set using uncorrected linear prediction estimates of natural sentences, the other using estimates made by hand. Phoneme restrictions imposed on linguistic properties allowed comparisons between continuous and intermittent voicing, oral or nasal and fricative manner, and unrestricted phonemic variation. Intelligibility tests revealed uniformly good performance with sentences created by hand-estimation and a minimal decrease in intelligibility with estimation by linear prediction due to manner variation with continuous voicing. Poorer performance was observed when linear prediction estimates were used to produce synthetic versions of phonemically unrestricted sentences, but no similar decline was observed with synthetic sentences produced by hand estimation. The results show a substantial intelligibility cost of reliance on uncorrected linear prediction estimates when phonemic variation approaches natural incidence. © 2011 Acoustical Society of America. [DOI: 10.1121/1.3631667]

PACS number(s): 43.71.Gv, 43.72.Ar, 43.72.Ja [MAH]

Pages: 2173–2178

I. INTRODUCTION

How accurate are the numerical methods of estimation used to analyze speech? A variety of approaches to automatic estimation is evident in the technical literature, linear prediction prominent among them. The limitations of this method are well known, and expositions of the hypothetical advantages and disadvantages appear in popular texts (for example, [Johnson, 1997](#)). In practice, there are relatively few reports that calibrate the efficacy of linear prediction in comparison to the unquestionably more laborious method of estimation by hand (for example, [Assmann and Katz, 2005](#); [Nearey et al., 2002](#); [Robb and Cacace, 1995](#)). Here, new measures comparing linear prediction and manual estimation are reported, motivated by an empirical aim to estimate the magnitude of the errors that accumulate and by a practical aim to appraise the potential of linear prediction for creating synthesis parameters in the production of sine-wave speech.

II. ESTIMATION BY LINEAR PREDICTION AND BY HAND

The acoustic effects of speech are regular, a consequence of the syllable cycle that produces an alternating rise and fall in energy radiated from the lips. The acoustic effects are also unpredictable, a consequence of the changing configuration of articulators that alters the shape of the resonating column of air excited by glottal, aspirated, and fricative

sources. Description of speech signals typically includes temporal measures, such as the duration and phasing of events portrayed in a time-domain representation, and spectral measures, such as the frequency of resonances portrayed in a frequency-domain representation. Linear prediction offers a method of estimating the frequency, amplitude, and bandwidth of vocal resonances, using an idealization of the speech spectrum and a statistical device to resolve the changes in frequency of the resonances occurring from moment to moment ([Markel and Gray, 1976](#)).

The idealization of the speech spectrum in linear prediction is simple to state. It assumes that the spectrum is vocalic with a fixed number of broad spectral peaks. The specific spectral model is set globally in an analysis, indifferent to momentary departures from idealization that are likely to occur in an acoustically varied speech sample. The idealization also assumes that the shape of the spectrum is negligibly affected by anti-resonances. Moreover, the analysis of each momentary sample is determined in part by preceding samples, an implicit assumption that the spectrum of speech changes gradually. These assumptions are the source of the advantage of linear prediction, computationally, and its disadvantages, analytically. For example, linear prediction representations of speech are likely to err when the actual spectrum differs from the idealization required by the model, whether by departing from the uniform harmonics characteristic of periodicity in glottal pulsing or in the number and frequency separation of broadband peaks. Errors in estimating the momentary shape of the spectrum are likely when there is a significant anti-resonance as typically occurs in the spectrum of a nasal or fricative. Estimates are also likely to

^{a)}Author to whom correspondence should be addressed. Electronic mail: remez@columbia.edu

be erroneous when the spectrum changes rapidly as occurs at moments of abrupt transition associated with the occurrence of nasals, fricatives, and the release of stop holds. Indeed, the model of speech as a fixed set of spectral peaks obliges the analysis to find peaks even during silent portions of speech, for instance, during the hold portions of voiceless stop consonants. These known liabilities of the analytical method impose a burden for the user of correcting the numerically obtained estimates at junctures likely to result in false values.

The project reported here sought to calibrate the accuracy of linear prediction estimates through the use of speech synthesis for empirical and practical aim alike. Empirically, the likely hazards of linear prediction estimation are well identified in principle, but the cost of erroneous estimation has rarely been assessed. For instance, errors should be less likely when speech is composed solely of phonemes that impose continuous and gradual change in the spectrum, for example, vowels and liquid consonants. Erroneous estimates should be relatively more likely with utterances composed of vowels, liquids, and nasals or composed of vowels, liquids, and voiced fricatives. Utterances composed of vowels, liquids, nasals, and voiceless stops or of vowels, liquids, nasals, and voiceless fricatives should be a greater challenge to linear prediction estimation due to the variation in the source of excitation. And, the likelihood of erroneous estimation by linear prediction should be greatest when the phonemic composition of speech is unconstrained. Yet, even if this distribution of errors is predictable, the cost of using an error-prone technique is useful to know. Regardless of the distribution, if errors are few and minor, the bench practice of using uncorrected linear prediction estimates need not change much. On the other hand, if the errors prove to be both numerous and substantial, research methods must adapt to this threat or suffer the effects of invalid estimates.

Practically, the laborious method of hand estimation used in the creation of sine-wave speech (Remez, 2008) imposes a heavy tax on the production of acoustic test materials for experiments and in the development of new projects. It will be useful to know whether the application of uncorrected linear prediction analysis in the production of sine-wave speech is practicable as some recent projects have presumed (Brungart *et al.*, 2006; Brungart *et al.*, 2005; Kamachi *et al.*, 2003; Möttönen *et al.*, 2006; Nitttrouer *et al.*, 2009; Vroomen and Baart, 2009).

III. METHOD

A. Acoustic test materials

Two intelligibility tests were created consisting of sine-wave sentences. One was synthesized from numerical estimates of spectral peaks created by linear prediction analysis of natural speech. The other was synthesized from estimates made by hand-tracing speech spectra displayed on a monitor. A single set of 70 natural utterances served as models for each kind of synthesis. There were 10 English sentences in seven types distinguished by phoneme composition: (1) vowels and liquid consonants (example: Will your lawyer allow our error?); (2) vowels, liquid, and nasal consonants

(example: I normally iron all morning.); (3) vowels, liquids, and voiced fricatives (example: Our rival loves the zoo.); (4) vowels, liquids, nasals, and voiceless fricatives (example: He swore he fell on a shoe in your house.); (5) vowels, liquids, nasals, and voiced stop consonants (example: Blend your red and blue dye.); (6) vowels, liquids, nasals, and voiceless stop consonants (example: In winter I park my car near town.); and (7) no compositional restrictions (example: The bark of the pine tree was shiny and dark.) Some sentences were original to this project, and some were taken from lists published by Huggins and Nickerson (1985), Kalikow, Stevens and Elliott (1977), Stubbs and Summerfield (1990) and the IEEE (1969) set, also known as the “Harvard Sentences.” A list of the sentences used in the study appears in the Appendix.

The natural speech samples on which the synthesis was based were spoken by one of the authors (R.E.R., an adult male) wearing a head-mounted microphone and seated in a sound attenuating chamber. The speech was sampled to disk at a rate of 44.1 kHz, and the sampled data were edited to produce 70 sentence-size files equated for loudness. Linear prediction sine-wave synthesis for each sentence was produced with PRAAT software (Boersma and Weenink, 2005) and a script written by Christopher J. Darwin that takes the peaks picked by the linear prediction analysis to set synthesis parameters for time-varying sinusoids. The synthesis parameters were used without correction. The script can be found at this location: http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts/SWS.

Estimates of frequency and amplitude of acoustic resonances, bursts, frictions, and murmurs were also produced by hand. Working in small groups, the seven co-authors inspected a spectrogram and waveform of a sentence displayed on a video monitor and used software to trace spectral prominences with the computer cursor. A trail of dots marked the traced paths of the selected spectral features. Each individual performing an analysis by hand identified the principal features by direct inspection of the spectrographic display and followed no formal set of procedures or rules. Frequency and amplitude were derived from the speech spectra for traced values at a temporal grain of 10 ms and were used to compile a table of synthesis parameters for each sentence. After an initial selection of spectral prominences occurred, interactive error correction ensued in which parameters were converted to sound and inaccuracy in the analysis was noted by listening critically. A varying number of cycles of synthesis and repair of the analysis occurred until the parameters were considered satisfactory. A comparison of the results of linear prediction and hand estimation for the same sentence is shown in the two panels of Fig. 1.

The sinusoidal complexes were converted to digital waveforms calculated with 16-bit amplitude resolution at a rate of 44.1 kHz and were stored in sampled-data format (Rubin, 1980). For the intelligibility tests, sentence sequences were transferred with no loss without resampling to optical disk, and were presented at a nominal level of 68 dB sound pressure level (SPL) via Beyerdynamic DT770 headphones to listeners seated in a sound-attenuating chamber.

FIG. 1. A comparison of estimation of formant patterns by linear prediction and by hand. Each panel shows a spectrogram of a natural utterance, “In winter I park my car near town,” with the estimated formant centers overlaid in white. (A) Estimates produced automatically by linear prediction. (B) Estimates produced by hand.

B. Participants

Forty listeners were recruited from the undergraduate population of Columbia University. Each reported no history of speech, language or hearing problem. None had prior exposure to sine-wave speech. A listener was assigned randomly to one of four test conditions.

C. Procedures

To create test sessions of moderate duration, sentence types were grouped into two sets. Twenty listeners, 10 who heard synthesis produced from linear prediction estimates and 10 from estimates produced by hand, were tested in Set A with sentences of these types: (1) vowels and liquids; (2) vowels, liquids, and nasals; (3) vowels, liquids, and voiced fricatives, and (7) no compositional restrictions. Twenty other listeners were tested in Set B, 10 with synthesis from linear prediction estimates and 10 with synthesis from hand estimates using sentences of these types: (4) vowels, liquids, nasals, and voiceless fricatives; (5) vowels, liquids, nasals, and voiced stop consonants; (6) vowels, liquids, nasals, and voiceless stop consonants; and (7) no compositional restrictions. Intelligibility was estimated from a transcription of each sentence written by a listener in a test booklet. A test session began with a warm-up sequence of six sine-wave sentences produced by the method of parameter estimation that matched the items of the main test block. In the warm up, the first three sentences in the series were transcribed for the listener in advance. The main test block was composed of 40 sentences, 10 of each type, presented in random order. Following the main test block was a six sentence cool-down series, a repetition of the warm-up except that none of the

sentences was transcribed for the listener. On each trial, a sentence was repeated five times with 2 s between repetitions. A listener was permitted to write while a sentence was playing.

IV. RESULTS

Each listener contributed 40 values to the dataset, the percent of syllables transcribed correctly on each sentence of the main test block. For the sentences in Set A, a repeated measures analysis of variance (ANOVA) was performed on the measures modeled on four levels of the within-subjects-factor phoneme composition and two levels of the between-subjects-factor estimation method. The analysis revealed a main effect of each factor, attributable to the better performance on some sentence types than others collapsing over estimation method and to better performance on hand-estimated sentences collapsing over sentence type. The statistics are: for phoneme composition, $F(3,16)=4.581$, $P<0.01$; for estimation method, $F(1,18)=1370.8$, $P<0.001$. At a finer grain, the analysis of variance also revealed an interaction of phoneme composition and synthesis method in transcription performance [$F(3,16)=7.833$, $P<0.002$]. The 95% confidence interval showed that the interaction in Set A was due solely to a difference in performance on sentences of unrestricted phonemic composition in each synthesis method. Performance did not differ between methods of spectral analysis when sentences were composed of vowels and liquids or of vowels, liquids, and nasals, or of vowels, liquids, and voiced fricatives. The results are shown in Fig. 2(A).

For the sentences in Set B, a repeated-measures ANOVA was performed on the transcription measures modeled on four levels of the within-subjects-factor phoneme composition and two levels of the between-subjects-factor estimation method. The analysis again revealed a main effect of each factor, attributable to the better performance on some sentence types than others collapsing over estimation method, and to better performance on hand estimated sentences collapsing over sentence type. The statistics are: for phoneme composition, $F(3,16)=9.904$, $P<0.001$; for estimation method, $F(1,18)=877.08$, $P<0.001$. At finer grain, the analysis also revealed an interaction of phoneme composition and method of estimation [$F(3,16)=14.417$, $P<0.001$]. The 95% confidence interval showed that the interaction in Set B was due to a difference in performance in two conditions, one in which sentences contained voiceless stop consonants and one in which phoneme composition was unrestricted. The results are shown in Fig. 2(B).

V. DISCUSSION

The outcomes of these tests show that the intelligibility of sine-wave speech modeled on natural utterances can be compromised when synthesis parameters are taken from uncorrected linear prediction estimates. Despite the known theoretical limitations of an all-pole vocalic model of speech spectra, the drastic reduction in detail exhibited in the sine-wave method is apparently forgiving of whatever errors in estimation might occur in representing the short-term

FIG. 2. Results of tests of intelligibility with sine-wave speech produced from two kinds of spectral estimation: Uncorrected linear prediction (unfilled bars) and estimation by hand (filled bars). (A) Results of tests with sentences composed of vowels and liquid consonants [V, L], vowels, liquid and nasal consonants [V, L, N], vowels, liquid and voiced fricative consonants [V, L, F(+V)], and unrestricted in phonemic composition. (B) Results of tests with sentences composed of vowels, liquid, nasal, and voiceless fricative consonants [V, L, N, F(-V)], vowels, liquid, nasal, and voiced stop consonants [V, L, N, S(+V)], vowels, liquid, nasal, and voiceless stop consonants [V, L, N, S(-V)], and unrestricted in phonemic composition. Each bar represents the average transcription of ten sentences provided by ten listeners. Error bars represent the 95% confidence estimate.

spectral shape of nasal and fricative segments. The evidence is seen in the similarity of good performance that was observed when sentences here were composed of nasal and voiced fricative consonants.

The rapid restructuring of the spectrum that occurs when the vocal source changed from voiced to voiceless posed a greater challenge as observed in the significant performance level difference when sentences contained voiceless stop consonants and when phoneme variation was not restricted. It is useful to note that this decrement in intelligibility was observed solely with synthesis produced from linear prediction estimates. Sine-wave sentences synthesized from hand-traced spectral peaks exhibited no similar decline in intelligibility as a consequence of the variety of phoneme types composing the test items.

Would a test based on consonant-vowel (CV) syllables have been more effective than one using sentences in showing the costs of erroneous estimation? The test sentences that were used here were semantically normal if unpredictable.

However, it is instructive to consider the role that sentence context might play in enhancing intelligibility as a consequence of the fluency of the speech and the convergence of constraint as a sentence progresses. On this topic, there is no reported direct comparison of sine-wave CVs derived from estimation by linear prediction and from estimation by hand. However, the identification of sine-wave CVs estimated by hand in an 18-alternative forced choice procedure was virtually error free when listeners were used to the unnatural timbre (Remez, 2008). This suggests that some segmental aspects of intelligibility of synthesis produced from accurate estimation can depend very little on sentence context. More generally, the intelligibility of segments and of sentences produced in the same way are typically correlated (Boothroyd and Nitttrouer, 1988), and for this reason, the value added by context can be understood as a straightforward consequence of segmental intelligibility. From this perspective, the benefits of sentence context in the counter-balanced tests reported here applied equally to each kind of synthesis. In these symmetrical conditions, the performance level differences diverged nonetheless under expectable conditions of phonemic restriction considering the errors of estimation that are likely with linear prediction. For these reasons, it is doubtful that a direct test using CVs would produce very different outcomes although definitive evidence on this matter could be produced by new tests.

Although this project revealed an intelligibility difference between hand estimation and uncorrected linear prediction estimation, the use of iterative cycles of analysis and correction has not been limited to hand estimation. The initial method for creating sine-wave speech relied on linear prediction estimates of natural speech followed by extensive hand correction (Remez *et al.*, 1981), and this method is effective in producing satisfactory test material (Roberts *et al.*, 2010). Other methods have included inspection of a time-aligned display of the waveform and the discrete Fourier analysis, using the linear prediction estimates to guide manual selection of spectral peaks (for example, Remez *et al.*, 1997). Whichever method is applied in the initial derivation of synthesis parameters, the intelligibility of the resulting synthesis evidently depends on correcting the errors.

It is constructive to note, too, that the perceptual consequences of erroneous estimation and the resulting deterioration in the quality of synthesis were observed here as performance level effects on intelligibility. This intelligibility difference prompts a suspicion of the likelihood of functional differences in the perception of sentences created by each method. Prior studies had found that synthetic speech and natural speech of nearly equivalent intelligibility differed nonetheless in cognitive effects. For example, memory tests that used both free and ordered recall of individual words presented in a list found poorer performance with synthetic speech relative to natural speech (Luce *et al.*, 1983). Even natural samples of CV syllables that had been spliced and recomposed electronically differed from unedited items in a speeded identification task (Whalen, 1984). This effect was plausibly attributed to conflicting acoustic correlates of segmental contrasts in the edited syllables that left the

identity of the segments unaltered. If such functional differences are evident when intelligibility hardly differs, it is reasonable to infer that intelligibility differences of the magnitude reported here are accompanied by differences in perceptual and cognitive function, as well.

While several methods exist for estimating the acoustic spectra of sampled speech, user-friendly linear prediction applications are available now at low cost or no cost. It is clear that this method of assessing the acoustic properties of speech is well established although the liabilities of the method and the need for correction of errors is less evident. Indeed, the distribution and size of errors of estimation are not often noted in the technical literature. To provide one index, this project aimed to assess the likely relation between phonemic composition and losses in intelligibility when synthetic speech is produced from linear prediction spectral estimates without correction. Accordingly, additional methods will be required to determine whether a straightforward means to improve synthesis from such estimates can be easily constructed. One key to understanding the potential for reducing the error-proneness of this analytical method is to determine whether the assumptions of the model oblige it to miss perceptually critical spectral details or, alternatively, to introduce spurious peaks at perceptually critical junctures. It might be possible to devise adaptive methods that relieve both of these liabilities of numerical estimation.

Nonetheless, on the basis of the present findings, it is potentially problematic to employ uncorrected linear prediction estimates as synthesis parameters when the phonemic composition of sentences is unrestricted or includes voiceless stop consonants. If scientific conditions require high levels of intelligibility without extensive training in an open-set response format that approaches natural conditions, these findings strongly encourage the correction of linear prediction estimates as synthesis parameters (see [Roberts et al., 2010](#)) and more broadly for characterizing the acoustic properties of speech.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the advice and helpful criticism of Robin Broder, Philip Rubin, and Michael Studdert-Kennedy. This research was supported by a grant from the National Institute on Deafness and Other Communication Disorders (DC-000308).

APPENDIX: SENTENCES OF VARYING PHONEME COMPOSITION USED IN INTELLIGIBILITY TESTS

1. Vowels and liquid consonants

I owe you a yoyo.
A war ally will rule Iowa.
Where were you well?
Larry wore a laurel a year early.
Lower your arrow Ella.
Why are you weary?¹
Will your lawyer allow our error?
I worry while you are away.

Are you aware I will roll away?
We all wear a rare yellow wool.

2. Vowels, liquid and nasal consonants

I am well known among men.¹
I normally iron all morning.
A royal memorial will remain.³
Are you a loyal union man?³
I owe no one any money.¹
You lie in an alarming manner.¹
I will marry you in May.³
I am wearing my maroon one.¹
When will our yellow lion roar?¹
You were wrong all along.¹

3. Vowels, liquid and voiced fricatives

Will they allow you a lawyer?
Lower the level of the revolver.
The weather is usually lovely.
Our rival will arrive early.
Is Lily loyal or evil?
Reserve the olives of the rural villa.
While I weigh the leather you use the razor.
Liver is always vile.
Our rival loves the zoo.
Will you reveal the loser of the war?

4. Vowels, liquid, nasal and voiceless fricative consonants

He swore he fell on a shoe in your house.
All mice seem similar from far away.
She will sell a flower near a sea shore.
A rainy Fall will follow a sunny Summer.
Will you seriously follow her?
How will you see a show for free?
If you wear a high heel shoe, you will suffer.
If you sin, you will learn your lesson.
A frail woman will feel safe on her sofa.
Soon our chef will weigh some flour.

5. Vowels, liquid, nasal and voiced stop consonants

An aid will guide you around our building.
A big wall would be a good border.
Bobby did a good deed.¹
Bend a band aid around your bloody elbow.
I needed a brand new rubber band.
Are you bored by your own name?
Do you abide by your bid?¹
A greedy boy died.¹
A deer and a bear will gladly dawdle in your garden.
Blend your red and blue dye.

6. Vowels, liquid, nasal and voiceless stop consonants

A parrot in a crate will talk all night.
I cannot tell a tale too well.
Our turtle ate a tiny kiwi.
In winter I park my car near town.

Can we keep your kite until tomorrow?
 I can knit one mitten per minute.
 Take a copy to Pete.¹
 Tell my uncle not to take our apple pie.
 We met you on time at an airport terminal.
 Can you write a term paper in a week?

7. Unrestricted phoneme composition

The beauty of the view stunned the young boy.⁴
 The steady drip is worse than a drenching rain.⁴
 The bark of the pine tree was shiny and dark.⁴
 The drowning man let out a yell.²
 They took the axe and the saw to the forest.⁴
 The boy was there when the sun rose.⁴
 Her purse was full of useless trash.⁴
 The sandal has a broken strap.²
 Two blue fish swam in the tank.⁴
 A pencil with black lead writes best.⁴

¹Huggins and Nickerson (1985).

²Kalikow, Stevens, and Elliott (1977).

³Stubbs and Summerfield (1990).

⁴IEEE (1969).

- Assmann, P. F., and Katz, W. F. (2005). "Synthesis fidelity and time-varying spectral change in vowels," *J. Acoust. Soc. Am.* **117**, 886–895.
- Boersma, P., and Weenink, D. (2005). "PRAAT: doing phonetics by computer (Version 4.3-01)." Retrieved from <http://www.fon.hum.uva.nl/praat/> (Last viewed 01/02/11).
- Boothroyd, A., and Nittrouer, S. (1988). "Mathematical treatment of context effects in phoneme and word recognition," *J. Acoust. Soc. Am.* **84**, 101–114.
- Brungart, D. S., Iyer, N., and Simpson, B. D. (2006). "Monaural speech segregation using synthetic speech signals," *J. Acoust. Soc. Am.* **119**, 2327–2333.
- Brungart, D. S., Simpson, B. D., Darwin, C. J., Arbogast, T. L., and Kidd, G., Jr. (2005). "Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task," *J. Acoust. Soc. Am.* **117**, 292–304.
- Huggins, A. W. F., and Nickerson, R. S. (1985). "Speech quality evaluation using phoneme-specific sentences," *J. Acoust. Soc. Am.* **77**, 1896–1906.
- IEEE (1969). "IEEE recommended practice for speech quality measurements," *IEEE Trans. Aud. Electroacoust.* **AU-17**, 225–246.

- Johnson, K. R. (1997). *Acoustic and Auditory Phonetics* (Blackwell, Cambridge, MA), pp. 3–165.
- Kalikow, D. N., Stevens, K. N., and Elliot, L. L. (1977). "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *J. Acoust. Soc. Am.* **61**, 1337–1351.
- Kamachi, M., Hill, H., Lander, K., and Vatikiotis-Bateson, E. (2003). "Putting the face to the voice: Matching identity across modality," *Curr. Biol.* **13**, 1709–1714.
- Luce, P. A., Feustel, T. C., and Pisoni, D. B. (1983). "Capacity demands in short-term memory for synthetic and natural speech," *Hum. Fact.* **25**, 17–32.
- Markel, J. D., and Gray, H. H., Jr. (1976). *Linear Prediction of Speech* (Springer, New York).
- Möttönen, R., Calvert, G. A., Jääskeläinen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., and Sams, M. (2006). "Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus," *NeuroImage* **30**, 563–569.
- Nearey, T. M., Assmann, P. F., and Hillenbrand, J. M. (2002). "Evaluation of a strategy for automatic formant tracking," *J. Acoust. Soc. Am.* **112**, 2323.
- Nittrouer, S., Lowenstein, J. H., and Packer, R. R. (2009). "Children discover the spectral skeletons in their native language before the amplitude envelopes," *J. Exp. Psychol. Hum. Percept. Perform.* **35**, 1245–1253.
- Remez, R. E. (2008). "Sine-wave speech," in *Encyclopedia of Computational Neuroscience*, edited by E. M. Izhikovich (cited as Scholarpedia, **3**, 2394).
- Remez, R. E., Fellowes, J. M., and Rubin, P. E. (1997). "Talker identification from phonetic information," *J. Exp. Psychol. Hum. Percept. Perform.* **23**, 651–666.
- Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (1981). "Speech perception without traditional speech cues," *Science* **212**, 947–950.
- Robb, M. P., and Cacace, A. T. (1995). "Estimation of formant frequencies in infant cry," *Int. J. Pediatr. Otorhin.* **32**, 57–67.
- Roberts, B., Summers, R. J., and Bailey, P. J. (2010). "The perceptual organization of sine-wave speech under competitive conditions," *J. Acoust. Soc. Am.* **128**, 804–817.
- Rubin, P. E. (1980). *Sinewave synthesis* {Internal memorandum} (pp. 1–20). New Haven, CT: Haskins Laboratories.
- Stubbs, R. J., and Summerfield, Q. (1990). "Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.* **87**, 359–372.
- Vroomen, J., and Baart, M. (2009). "Phonetic recalibration only occurs in speech mode," *Cognition* **110**, 254–259.
- Whalen, D. H. (1984). "Subcategorical mismatches slow phonetic judgments," *Percept. Psychophys.* **35**, 49–64.