

Asynchrony tolerance in the perceptual organization of speech

ROBERT E. REMEZ, DARIA F. FERRO, STEPHANIE C. WISSIG, AND CLAIRE A. LANDAU
Barnard College, New York, New York

Researchers have claimed that listeners tolerate large temporal distortion when integrating the spectral components of speech. In some estimates, perceivers resolve linguistic attributes at spectral desynchronies as great as the duration of a syllable. We obtained new measures of perceptual tolerance of auditory asynchrony, using sine-wave synthesis in order to require perceivers to resolve the speech stream dynamically. Listeners transcribed sentences in which the tone analogue of a second formant was desynchronized relative to the remaining tones of a sentence, with desynchrony ranging from a 250-msec lead to a 250-msec lag. Intelligibility declined symmetrically from 72% at synchrony to 7% at ± 100 msec. This finding of narrow asynchrony tolerance indicates a time-critical feature of the auditory perceptual organization of speech.

How does a listener resolve disparate acoustic elements into a single perceptual stream? The evident coherence and the continuity of speech are a challenge to accounts of perceptual organization. Vocally produced sound is distributed across seven octaves, and the perceptual coherence of a speech stream therefore entails integration of spectral attributes spread over a wide frequency range. Additionally, the acoustic variety of the constituents of speech requires the perceptual coherence of dissimilar and discontinuous elements. The perceptual aggregation of whistles, clicks, hisses, buzzes, and hums in a speech stream almost certainly forms as a consequence of their coordinate variation (Remez, 2005; Remez, Rubin, Berns, Pardo, & Lang, 1994). Relying on sensitivity to dynamic coherence, listeners can follow a speech stream even when its spectral elements are neither physically nor subjectively speech-like. Such dynamic sensitivity without the contribution of natural acoustic vocal products is a likely cause of the perceptual coherence of sine-wave speech (Remez, Rubin, Pisoni, & Carrell, 1981; Rubin, 1980), noise-band vocoded speech (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995), and acoustic chimeras of speech (Smith, Delgutte, & Oxenham, 2002).

In contrast to dynamic sensitivity to coordinate variation, in element-based accounts, researchers have proposed that auditory perceptual integration relies chiefly on sensitivity to local similarity among detailed spectral elements. Explanations of this kind offer an appealingly simple description of stream formation applicable to simple sources of sound (Bregman, 1990; Darwin, 1997) but fall short of explaining the perceptual organization of complex streams, including speech. Despite the claim that speech perception requires special cognitive functions (Lieberman & Mattingly, 1989), complex sound streams

are not unique to speech, and no evidence presently implicates a specialized organizational function devoted to spoken sources (Remez et al., 1994). The characteristics of dynamic sensitivity, though, can be productively examined in the case of speech, because this sound source is acoustically well modeled and much is known about the psychoacoustics of these sounds.

Temporal Integration and Asynchrony Tolerance in Speech

Researchers have also claimed that dynamic sensitivity in the perceptual organization of speech is necessary for the temporal integration of sensory components, whether the presentation is auditory (Saberri & Perrott, 1999) or audiovisual (Munhall, Gribble, Sacco, & Ward, 1996). However, some key auditory experiments have provided evidence about the parameters of dynamic sensitivity that is uncertain for two reasons: First, the perceptual effects of dynamic sensitivity were not assessed in strict independence from natural acoustic spectral details. Second, the performance of listeners varied over a narrow range, making any perceptual effects difficult to resolve in the behavioral measures. For example, Greenberg and Arai (1998) used a speech spectrum divided into 19 $\frac{1}{4}$ -octave pass bands that were systematically temporally misaligned over a range of 60–240 msec. The desynchronization of adjacent frequency bands was controlled in order to preclude local pockets of synchrony. Intelligibility was unaffected at desynchronies < 80 msec and was minimally compromised at desynchronies < 140 msec (cf. Fu & Galvin, 2001). Even at the extreme of 240 msec, intelligibility had fallen only to 50%. It might seem that this finding is consistent with claims that perceptual integration is keyed to the pace of syllables, within the range of 3–8 Hz (Green-

R. E. Remez, remez@columbia.edu

berg, 1999; Saberi & Perrott, 1999; cf. Studdert-Kennedy, 1976). However, the use of filtered natural speech necessarily preserves fine-grained natural acoustic vocal products and allows their perceptual effects, complicating the interpretation of the finding.

In related tests, Silipo, Greenberg, and Arai (1999) also desynchronized narrow-frequency bands of filtered natural speech, dividing the spectrum into 14 $\frac{1}{3}$ -octave pass bands. Of those 14, only 4 widely spaced bands were kept to compose the test material, and the remaining 10 bands were discarded, resulting in a sparse signal of four narrow spectral slits. The bands had these cuts: Band 1, 298–375 Hz; Band 2, 750–945 Hz; Band 3, 1890–2381 Hz; and Band 4, 4762–6000 Hz. Asynchronies of 25–75 msec were imposed parametrically on one or more bands, and under these conditions an asynchrony of 25 msec resulted in intelligibility decreased by 10%, relative to a synchronized control of 88%. However, with the exception of the pairing of the lowest and the highest band, intelligibility at the greatest asynchrony of 75 msec remained good, at 56% for one and 41% for two desynchronized pass bands. Such performance is likely to reflect both the integration of the four spectral bands despite desynchrony and, when integration failed, the compensation by listeners on the basis of the richness of the remaining spectrum. A more discriminating test is required to identify perceptual integration, using intelligibility as the measure.

In order for a test to offer a sensitive measure of perceptual coherence of spectrotemporal variation, critical conditions must be met. Each item in the test must be composed of acoustic ingredients that evoke clear linguistic impressions when they are presented together, yet each ingredient presented alone must not evoke linguistic impressions. To isolate the perceptual effects of the individual elements from the effects of the pattern that they compose, studies have also used intelligible spectral patterns composed of elements that are unlike speech in detail (Remez et al., 1981; Shannon et al., 1995; Smith et al., 2002). If these conditions can be satisfied, an intelligibility measure can be used to probe for perceptual integration of the acoustic ingredients.

In the present study, we aimed to satisfy this objective by using sine-wave replicas of sentences (Remez et al., 1994; Remez et al., 1981). These highly intelligible items necessarily draw perception away from sensory details, requiring attention to dynamic spectrotemporal attributes for integration. In prior studies, researchers have revealed that intelligibility of sine-wave speech is lost when the tone analogue of the second formant is absent (Remez et al., 1994; Remez et al., 1981). For this reason, transcription performance can be used as a measure of perceptual integration when this critical tone is present but desynchronized from the rest of the sine-wave pattern. In addition, by relying on tone analogues of speech instead of on natural or synthetic speech, we aimed to achieve a clean measure of dynamic sensitivity in this study, independent of the auditory effects of acoustic constituents characteristic of natural speech.

EXPERIMENT

Desynchronizing the Tone Analogue of the Second Formant

Method

Acoustic test material. Fifty sentences, drawn from the phonemically balanced IEEE set (Egan, 1948) and the Speech Perception in Noise set (Kalikow, Stevens, & Elliott, 1977), were spoken by a male talker (R.E.R.) seated in a sound-attenuating chamber. The sentences were phonemically, lexically, and syntactically diverse, and were intended to be difficult to guess (see the Appendix). The speech was sampled at 22.05 kHz. Individual sentence samples were equated for amplitude and were analyzed in order to compose parameters for a sine-wave synthesizer. Frequency and amplitude of acoustic resonances, bursts, frictions, and murmurs were interactively estimated by tracing acoustic features of the natural samples presented in a spectrographic display, and the selected values were used to compile a synthesis table. The synthesis parameters represented frequency and amplitude values of four time-varying sinusoids at a grain of 10 msec throughout each utterance. The associated waveforms were calculated at a sampling rate of 44.1 kHz with 16-bit resolution of amplitude and were stored in sampled-data format (Rubin, 1980). A pilot test established the intelligibility of these 50 sentences; a mean of 93% of the syllables were identified correctly ($SD = 10.9$; max. = 100, min. = 39.7).

A second pilot test was conducted using the 50 sentences, with the tone analogue of the second formant deleted from them. This test provided an estimate of intelligibility under conditions in which the second formant tone was not available to be integrated with the other tonal components. A subset of 15 of the 50 sentences was selected for use as test items in asynchrony conditions; each of these sentences was intelligible when all tones were present, and largely unintelligible when the second formant analogue was absent. The intelligibility of second formant knock-out sentences was poor overall: 15.5% of the syllables were identified correctly ($SD = 12.9$; max. = 55.5, min. = 1). The 15 sentences selected for this procedure were identified in the two pilot tests at 98% of the syllables correct when they were intact, and 6% correct when the second formant analogue was removed. Basing the test of asynchrony tolerance on these 15 sentences provided a wide range of performance levels within which to see the effects of integration and its absence.

The properties of segmental contrasts do not uniformly correlate with frequency and amplitude variation across the formants. Variation in the first formant is often associated with contrasts in the manner and voicing of consonants and in the height of vowels, whereas variation in the second formant is associated with contrasts in the articulatory place of consonants and the advancement of vowels. This distinction is based on modeling of acoustic phonetics (Stevens, 1999); there are no acoustic norms for phonetic features, and therefore no firm empirical grounds to suspect an effect on asynchrony tolerance of phonetic properties due to their expression in narrower or wider temporal grain. Nonetheless, in a test in which the tone analogue of the second formant is desynchronized, a principal acoustic correlate of phonetic place and advancement is also plausibly desynchronized.

Each sentence was synthesized in its veridical temporal pattern, nominally the 0-msec asynchrony, and in departures from this pattern created by offsetting the second formant tone. The gradient of asynchrony used 50-msec steps from 50 msec through 250 msec of lead and from 50 msec through 250 msec of lag of the analogue of the second formant relative to the remainder of the sine-wave sentence pattern. Test items were stored in sampled-data format, were transferred to compact disc, and were played for listeners at the time of testing. They were presented at a nominal level of 68 dB SPL via Beyerdynamic DT770 headphones to listeners seated in a sound-attenuating chamber.

Participants. One hundred nine undergraduate volunteers from the population of Barnard College received credit toward a course

requirement for participating. Each was a native speaker of English and reported normal hearing at the time of testing. Listeners were tested in groups of 8 or fewer. A brief transcription pretest of six intact, novel sine-wave sentences was given, in order to assess the susceptibility of each listener to speech perception from time-varying sine waves. Eighty-nine listeners, who performed at 80% correct or better on the pretest, were included in the study. One listener was excluded from the experiment for failing to follow instructions, which left 88 listeners contributing to the data set.

Procedure. Eleven test sessions were created; each listener was randomly assigned to one of these. Within a session, 11 asynchronies occurred. The 15 sentences were rotated through asynchronies pseudorandomly. The order of asynchrony was random within a test session. A trial consisted of five repetitions of a sine-wave sentence, during which a listener was asked to transcribe it.

Results and Discussion

Each listener contributed 11 measures of transcription accuracy—the percentage of syllables transcribed correctly—one for each degree of asynchrony that occurred within a test block. The performance level at each asynchrony was determined on the basis of the transcriptions of 8 participants per test block across 11 blocks; group averages therefore represent the means of 88 points for each value of asynchrony. A one-way repeated measures ANOVA revealed a significant effect of asynchrony [$F(1,10) = 86.1, p < .001$].

The relation observed in this test between asynchrony of the second formant tone analogue and transcription is

simple to state. The linguistic attributes of the sentence were readily resolved when the tones were synchronized—that is, when they were temporally veridical; transcription performance was at 72% of the syllables transcribed correctly. At 50-msec departures from synchrony, transcription accuracy fell by more than half, as compared with the performance at synchrony, whether the second formant tone led (33% correct) or lagged (30% correct). Performance was exceedingly poor at asynchronies greater than 50 msec. The function relating performance to asynchrony is also symmetrical about the point of synchrony, in contrast to the findings in the audiovisual case, in which tolerance of asynchrony favored lagging over leading audible sensory samples of speech (Grant, van Wassenhove, & Poeppel, 2004; Munhall et al., 1996). Figure 1 shows the group results of this test. Each point represents the average accuracy in syllables transcribed correctly. The error bars represent the 95% confidence interval determined in the repeated measures analysis.

These measures of tolerance of auditory asynchrony in the perceptual integration of sine-wave speech differ from some of the reports that have provided a precedent for this study. Some prior findings, using digitally manipulated natural speech, have shown that perceivers integrate audible attributes of speech at asynchronies as great in duration as a syllable. Such wide tolerance in perceptual integration had motivated an explanation of integration

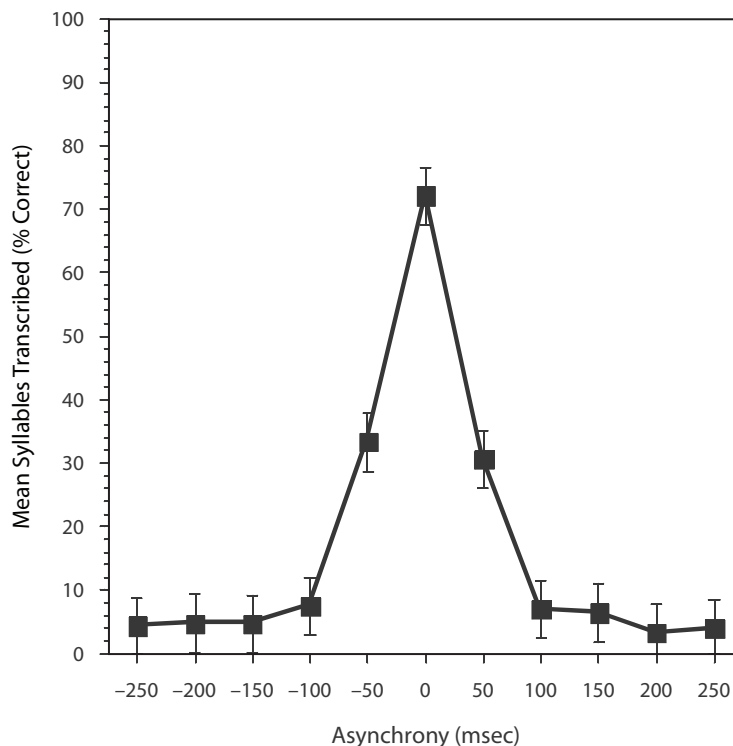


Figure 1. Group mean transcription performance shown as a function of the asynchrony of the tone analogue of the second formant. Asynchrony is the parameter of the *x*-axis, with the veridical temporal relation between the tone analogue of the second formant set to 0 msec, and variants departing in 50-msec steps from that value.

in speech perception invoking the intrinsically syllabic nature of speech production. However, if an integrative perceptual function is indexed by the precise temporal characteristics of its tolerance of asynchrony, then the integration of dynamic acoustic correlates of speech occurs within an interval far briefer than a syllable.¹

GENERAL DISCUSSION

In the present study, we aimed to clarify the characteristics of asynchrony tolerance in the perceptual organization of speech. Although a precise explanation requires additional tests, the present findings suggest that dynamic auditory sensitivity to linguistically governed vocalization does not exhibit wide asynchrony tolerance. In contrast to measures using digitally manipulated natural speech, the use of sine-wave replication eliminated the spectral details characteristic of natural speech and the auditory qualities that they evoke, compelling the perceiver to integrate the acoustic constituents largely on the basis of their coordinate spectrotemporal variation. Nonetheless, there are several cautions to consider before endorsing so general an interpretation of the findings of this study.

First, sine-wave speech is perceptually bistable (Remez, Pardo, Piorkowski, & Rubin, 2001). Despite the intelligibility of sine-wave sentences, generic functions of auditory organization split the tones into separate perceptual streams, while the coordinate variation in the tones promotes integration. These stable alternate organizations of sine-wave speech are concurrent, in contrast to the ambiguous figures, reversible-perspective drawings, or boundary-contour bistabilities in visual perception (see note 5 in Remez et al., 1994). It is possible that this organizational bistability blocked the action of integrative functions that entail asynchrony tolerance. However, the linguistic organization of sine-wave speech is readily apprehended and requires no special training, no long-term experience, and no induction of cognitive resources qualitatively different from those of natural speech. Indeed, the ease with which listeners adjust to the anomalous timbre has been taken as evidence of the ordinariness of the perceptual functions that they engage. Therefore, the bistability of these sounds is probably not responsible for the lack of tolerance of asynchrony.

Second, the absence of speech-like short-term spectra deprives sine-wave speech of the subjective quality of natural vocalization and, under these circumstances, there is little besides the dynamic properties of the time-varying tones to evoke an impression of speech. Accordingly, if tolerance of asynchrony depended on the familiar quality of speech, this function would be blocked by the sine-wave form of speech. Could the absence of natural vocal quality produce a great change in perceptual function, including the suspension of asynchrony tolerance? Prior studies have shown that the resolution of phonetic attributes of speech can be largely unaffected by extreme departures from natural vocal quality (Shannon et al., 1995; Smith et al., 2002), although it remains possible that such performance levels were achieved through heroic compensation on the part of listeners. This aspect of speech perception

is simply not well explored. It remains to be determined whether tolerance of auditory asynchrony in speech perception is similarly indifferent to timbre. Such a test will present a formidable technical challenge, though. In order to desynchronize a proper second formant without creating a desynchronized fundamental frequency contour as well, speech synthesis would require a constant fundamental frequency for all vocalic constituents. This would accomplish the objective of creating test items within which a formant band was progressively desynchronized without also desynchronizing the natural pitch contour of the original sentence. However, this method would also eliminate the natural quality attributable to typical variation in glottal period through a sentence, and would thereby oppose the aim of achieving a natural quality. A solution to this empirical problem that uses fluent speech or sentence-length samples is not obvious.

Third, it is possible that the design of the present test was, itself, responsible for the findings of such narrowly tuned asynchrony tolerance. Sentences for the test were chosen so that they were highly intelligible when the sine-wave components were integrated and barely intelligible when integration was impossible. These conditions surely imposed a steep decline in performance when integration was hampered. Because the desynchronized components were individually unintelligible and unpredictable from trial to trial, there was little opportunity for a participant in our test to act strategically.

One clue about this aspect of perception is available in the performance-level contrast between the same intact sentences presented in the pretest and in the asynchrony test. In the pretest, temporally veridical sine-wave sentences were presented for transcription as a series. In the asynchrony test, those identical items were presented among test items in which a key tone in each sentence exhibited a randomly different degree of desynchrony on each trial. The acoustically identical test items were intelligible at 98% in the pretest, but at 72% in the asynchrony test. This performance-level difference shows that changing the asynchrony unpredictably from trial to trial imposes a cost on perceptual integration. It also suggests the adaptability of dynamic sensitivity to time-varying properties if these are consistent over the short term. Some evidence of this kind of adaptability has been reported for audiovisual asynchrony tolerance (Navarra et al., 2005). New tests will be needed in order to obtain an understanding of the conditions in which the temporal window of perceptual integration is narrow, as reported here in a unimodal case, and those in which the temporal window of integration is wide.

AUTHOR NOTE

The authors are grateful for the advice and encouragement of Robin Broder, Morgana Davids, Kathryn Dubowski, Isabel Jay, Jennifer Pardo, and Philip Rubin. This research was supported by an award from the National Institute on Deafness and Other Communication Disorders (DC00308) to Barnard College. Correspondence concerning this article should be addressed to R. E. Remez, Department of Psychology, Barnard College, 3009 Broadway, New York, NY 10027-6598 (e-mail: remez@columbia.edu).

REFERENCES

- BREGMAN, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.
- DARWIN, C. J. (1997). Auditory grouping. *Trends in Cognitive Sciences*, **1**, 327-333.
- EGAN, J. P. (1948). Articulation testing methods. *Laryngoscope*, **58**, 955-991.
- FU, Q.-J., & GALVIN, J. J., III (2001). Recognition of spectrally asynchronous speech by normal-hearing listeners and Nucleus-22 cochlear implant users. *Journal of the Acoustical Society of America*, **109**, 1166-1172.
- GRANT, K. W., VAN WASSENHOVE, V., & POEPEL, D. (2004). Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony. *Speech Communication*, **44**, 43-53.
- GREENBERG, S. (1999). Speaking in shorthand—A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, **29**, 159-176.
- GREENBERG, S., & ARAI, T. (1998). Speech intelligibility is highly tolerant of cross-channel spectral asynchrony. In P. Kuhl & L. Crum (Eds.), *Proceedings of the Joint Meeting of the Acoustical Society of America and the International Congress on Acoustics* (pp. 2677-2678). Melville, NY: Acoustical Society of America.
- KALIKOW, D. N., STEVENS, K. N., & ELLIOTT, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America*, **61**, 1337-1351.
- LIBERMAN, A. M., & MATTINGLY, I. G. (1989). A specialization for speech perception. *Science*, **243**, 489-494.
- MUNHALL, K. G., GRIBBLE, P., SACCO, L., & WARD, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, **58**, 351-362.
- NAVARRA, J., VATAKIS, A., ZAMPINI, M., SOTO-FARACO, S., HUMPHREYS, W., & SPENCE, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research*, **25**, 499-507.
- REMEZ, R. E. (2005). Perceptual organization of speech. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 28-50). Oxford: Blackwell.
- REMEZ, R. E., PARDO, J. S., PIORKOWSKI, R. L., & RUBIN, P. E. (2001). On the bistability of sine wave analogues of speech. *Psychological Science*, **12**, 24-29.
- REMEZ, R. E., RUBIN, P. E., BERNS, S. M., PARDO, J. S., & LANG, J. M. (1994). On the perceptual organization of speech. *Psychological Review*, **101**, 129-156.
- REMEZ, R. E., RUBIN, P. E., PISONI, D. B., & CARRELL, T. D. (1981). Speech perception without traditional speech cues. *Science*, **212**, 947-949.
- RUBIN, P. E. (1980). *Sinewave synthesis* [Internal memorandum]. New Haven, CT: Haskins Laboratories.
- SABERI, K., & PERROTT, D. R. (1999). Cognitive restoration of reversed speech. *Nature*, **398**, 760.
- SHANNON, R. V., ZENG, F.-G., KAMATH, V., WYGONSKI, J., & EKELID, M. (1995). Speech recognition with primarily temporal cues. *Science*, **270**, 303-304.
- SILIPO, R., GREENBERG, S., & ARAI, T. (1999). Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations. *Proceedings of Eurospeech 1999* (pp. 2687-2690). Grenoble: European Speech Communication Association.
- SMITH, Z. M., DELGUTTE, B., & OXENHAM, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, **416**, 87-90.
- STEVENS, K. N. (1999). *Acoustic phonetics*. Cambridge, MA: MIT Press.
- STUDDERT-KENNEDY, M. (1976). Speech perception. In N. J. Lass (Ed.), *Contemporary issues in experimental phonetics* (pp. 243-293). New York: Academic Press.

NOTE

1. A second test was also constructed on this model, with a desynchronized tone analogue of the first formant. A pretest was conducted to calibrate the intelligibility of the sentences with and without the first formant tone, and to identify the test items offering the greatest performance range in which to observe the perceptual effects of asynchrony on perceptual integration. A new group of listeners, unfamiliar with the test items or protocol, was recruited, and the procedure for testing and evaluating the desynchrony of the first-formant tone was the same as the conditions using a desynchronized second-formant tone. The outcome of this test was much the same as is described here for the measures of a desynchronized second-formant tone. Performance was symmetrical about the point of synchrony and was poor at asynchronies greater than 50 msec, indicating the loss of perceptual integration of the components. Although it is arguable that classes or dimensions of phonetic attributes differ as to which spectral regions convey them effectively, a comparison of these results suggests that the auditory integration of the dynamic acoustic properties of speech is not contingent on synchronization of one or another dimension. Nor does it appear that asynchrony tolerance in the auditory modality favors a specific frequency band or phonetic dimension.

APPENDIX

The Sentences Used in the Experiment

A pencil with black lead writes best.
 Cut the meat into small chunks.
 Football is a dangerous sport.
 He ran halfway to the hardware store.
 Her purse was full of useless trash.
 His boss made him work like a slave.
 Press the pants, and sew a button on the vest.
 The bark of the pine tree was shiny and dark.
 The beauty of the view stunned the young boy.
 The bill was paid every third week.
 The drowning man let out a yell.
 The sandal has a broken strap.
 The steady drip is worse than a drenching rain.
 The watchdog gave a warning growl.
 Two blue fish swam in the tank.

(Manuscript received June 27, 2007;
 revision accepted for publication March 7, 2008.)