The intelligibility of pointillistic speech

**Gerald Kidd, Jr., Timothy M. Streeter, Antje Ihlefeld,
Ross K. Maddox, and Christine R. Mason**
Hearing Research Center, Boston University, Boston, Massachusetts 02215
gkidd@bu.edu, timstr@cns.bu.edu, antje1@gmail.com, rkmaddox@bu.edu, cmason@bu.edu

Abstract: A form of processed speech is described that is highly discriminable in a closed-set identification format. The processing renders speech into a set of sinusoidal pulses played synchronously across frequency. The processing and results from several experiments are described. The number and width of frequency analysis channels and tone-pulse duration were variables. In one condition, various proportions of the tones were randomly removed. The processed speech was remarkably resilient to these manipulations. This type of speech may be useful for examining multitalker listening situations in which a high degree of stimulus control is required.

1. Introduction

It has been demonstrated in a variety of ways that human speech is remarkably resilient and can convey meaning even under conditions of extreme distortion. For example, early work on "infinitely" peak-clipped speech revealed that the speech retained a high degree of intelligibility (e.g., Licklider and Pollack, 1948). It was also found that speech could be interrupted frequently and yet still be understood (Miller and Licklider, 1950). Furthermore, it has long been known that the information in speech is distributed across a wide range of frequencies and conveys meaning through the variation over time within these different frequency channels. Limiting the information to a subset of channels can provide some (highly predictable) degree of intelligibility (e.g., French and Steinberg, 1947). More recently interest in the essential aspects of speech has increased due to the development of cochlear implants. Shannon *et al.* (1995) described a means for simulating cochlear implant processing and demonstrated that such speech could be highly intelligible. In their procedure, the amplitude envelopes of several bands of speech are extracted and used to modulate noiseband carriers limiting the speech cues primarily to those conveyed by the envelopes. This type of "vocoded" speech has a long history (Dudley, 1939) and has been utilized and modified in various ways by many recent investigators (e.g., Dorman *et al.*, 1997; Loizou *et al.*, 1999; Arbogast *et al.*, 2002; Qin and Oxenham, 2003; Yang and Fu, 2005; Brungart *et al.*, 2005; Nie *et al.*, 2005; Poissant *et al.*, 2006; Throckmorton *et al.*, 2006; Stickney *et al.*, 2007; Whitmal *et al.*, 2007; Souza and Rosen, 2009).

A method of representing speech is presented here which combines some features of previous methods in addition to more severely quantizing the information. This speech is referred to as "pointillistic speech" because in the limit the speech signal is reduced to a time-frequency matrix of points with each point (or "element") consisting of a brief pulsed "pure" tone represented by only two values (its frequency and amplitude). Using this technique, speech identification results are reported providing a parametric examination of the effects of manipulating variables in the processing (number of time elements, number of frequency analysis channels, and proportion of resulting matrix removed).

2. Methods

2.1 The processing algorithm

The speech was filtered into 4, 8, or 16 contiguous frequency bands spaced logarithmically with a total range of $267-10667$ Hz. Within each analysis band, the Hilbert magnitude and phase were
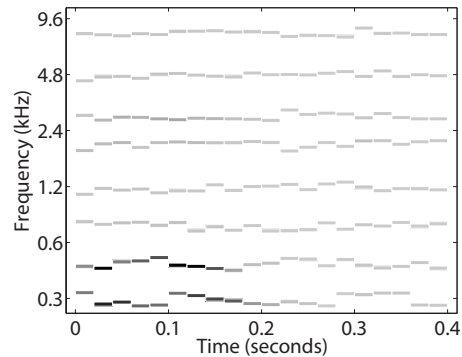
Fig. 1. A schematic illustration of the result of the processing in which the speech signal is replaced by a matrix of tone bursts. In this example there were eight tones per time window that were each 20 ms.

computed yielding functions describing the variation in the amplitude envelopes and instantaneous phases over time. The average values of the envelope and frequency were computed in each band for discrete contiguous time windows of 10, 20, 40, and 80 ms. The envelope value was calculated by squaring the mean of the absolute value of the Hilbert envelope and the frequency value was determined by taking the first time-derivative of the instantaneous phase function averaged over all positive frequencies. This yielded two numbers representing the stimulus at each time-frequency point. Then, a 0°-phase sinusoid of that amplitude and frequency at the total duration of the time window (including a 3-ms rise-decay) was created. The resulting sinusoidal elements were concatenated in time and summed across frequency producing sets of temporally contiguous, non-overlapping, synchronously gated tones. Figure 1 illustrates the result of this process for a single word under one of the conditions (8 tones, 20 ms) used in Exp. 1. The resulting speech has some of the envelope and fine structure information preserved, as in cochlear implant simulation (vocoded) speech (e.g., Nie *et al.*, 2005; Throckmorton *et al.*, 2006; Stickney *et al.* 2007), combined with more quantized envelope time segments (e.g., Loizou *et al.*, 1999; Brungart *et al.*, 2007; Li and Loizou, 2008).

2.2 Listeners

A total of 21 paid listeners (ages 19–31) participated in this study with 11 listeners participating in Exp. 1, 6 listeners participating in Exp. 2, and 6 listeners participating in Exp. 3. Two of the listeners participated in both Exps. 1 and 3 but none had participated in previous experiments in this laboratory.

2.3 Stimuli

The speech was from a laboratory-designed monosyllabic corpus (Kidd *et al.*, 2008) consisting of eight tokens from each of five categories: ⟨*subject*⟩ ⟨*verb*⟩ ⟨*number*⟩ ⟨*adjective*⟩ ⟨*object*⟩. In most conditions, for each trial, five words (one from each category without replacement) from one randomly chosen talker (of eight males and eight females) were concatenated in syntactically correct order. For one condition in Exp. 3, five of the entire 40 words were randomly chosen and concatenated on each trial producing sequences that were very unlikely to be syntactically correct.

2.4 Procedures

Listeners were seated in a sound-treated IAC chamber wearing Sennheiser HD280 Pro earphones. Stimuli were presented diotically through Tucker-Davis Technologies hardware at 60 dB sound pressure level. Listeners were instructed to report all five keywords by clicking on a response graphical user interface (GUI). The GUI had a button for each keyword organized in columns according to the word categories and within each column sorted in alphabetical (or in the case of numbers, numerical) order. Each keyword was scored individually and the listeners received no feedback. In Exps. 1 and 2, each listener was tested in 12 conditions: 4 time windows by 3 number-
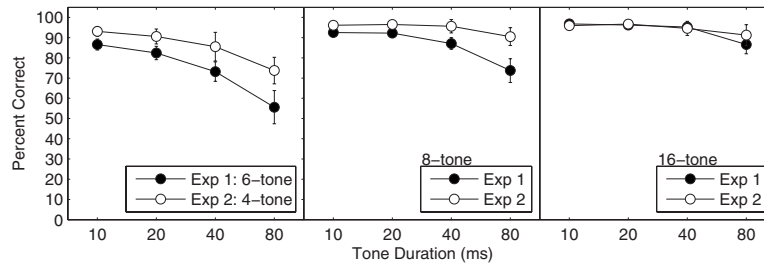
Fig. 2. Group mean word identification from Exps. 1 and 2. Error bars are ±1 standard deviation.

of-analysis-bands cases. All 12 conditions were presented three times in random order in each block of five-word trials and each listener completed 12 blocks. In Exp. 1, the tones for every band, every other band, or every third band of the 16-band case (starting with the lowest frequency) were used yielding 6, 8, or 16 simultaneous tones in each time window. For Exp. 2, the 16-tone case was identical and the 4- and 8-tone cases (in which the original analysis bands were wider) were also tested. In Exp. 3, only one combination of time windows and number of tones was tested (16 tones, 10 ms), while the number of elements in the signal was manipulated by randomly removing various proportions (0, 0.5, 0.66, 0.75, and 0.875) of the time-frequency bins representing the signal on a per-word per-trial basis. All proportions were tested for both syntactically correct and random order five-word utterances. Each block consisted of 15 trials for each of the five proportions in one of the word order conditions. Eight blocks were completed by each listener. Each experiment lasted about 2 hours.

3. Results

3.1 Experiments 1 and 2: Effect of number of analysis bands and duration of time windows

The results of Exps. 1 and 2 are shown in Fig. 2. Pointillistic speech, at least as assessed in this closed-set paradigm, is highly identifiable under certain combinations of stimulus parameters. Predictably, for any time window, performance was best for the highest number of tones and declined as that number decreased. Furthermore, for any given number of tones, performance generally decreased as the tone duration increased beyond 20 ms. The highest group-mean scores overall were obtained for the 16-tone signal at 10, 20, or 40 ms tone durations where the group mean values obtained were approximately 95% correct or better. However, even for the conditions with fewer tones, performance was better than 85% correct for the 10 ms duration. The poorest performance was for the fewest tones and longest time segments which provided the coarsest representation. However, even in that condition listeners achieved about 55% correct identification when 6 of the 16 analysis bands were included and almost 74% correct when the information was extracted from four wide bands covering the entire range. The variability across listeners was generally rather small but increased for the conditions under which performance was relatively poor. The 16-tone case allows a comparison of the two groups of listeners. Overall the results are remarkably similar with the 6 listeners in Exp. 2 performing slightly better than the 11 in Exp. 1 for the longest duration. In general, performance is better when the information comes from a wider analysis band (Exp. 2) than when narrow bands are used but some are excluded. In the left panel this is true even when the wide band analysis is only four tones as compared to six tones from every third band in the original set of 16.

3.2 Experiment 3: Effect of random removal of elements

The results are shown in Fig. 3 for the syntactically correct and random word order conditions. Rather remarkably the speech was highly robust with respect to this type of distortion. When the target words were presented in correct syntactic order, removing 2/3 of the elements only reduced identification performance to 92% correct. For comparison, the asterisks are results from Exp. 1 for the 10-ms six-tone and eight-tone cases. These correspond to proportions removed of
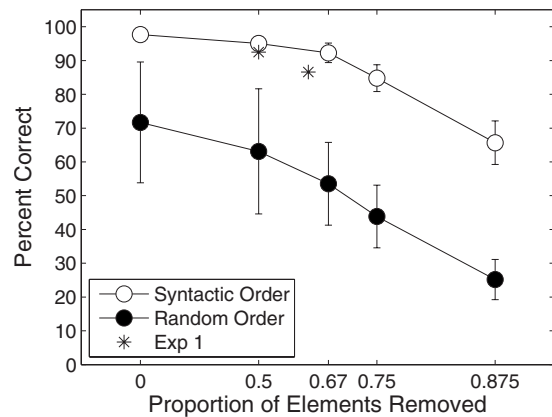
Fig. 3. Group mean word identification from Exp. 3. Error bars are ± 1 standard deviation.

0.625 and 0.5, respectively. As can be seen, performance is somewhat worse when the removed elements are from fixed frequency regions across all time. For words in random order, the effect of removing elements was greater, with mean performance of about 52% correct when only 1/3 of the elements remain dropping to about 25% correct when only 1/8 of the elements remain. The difference between syntactic and random order doubtless reflects the advantage of appropriate contextual order as well as some inherent differences in chance performance related to guessing or the ability to match the distorted token to 1 of 8 versus 1 of 40 alternatives. There were also larger differences across listeners for the random word order conditions. However, it is clear that these words may be discriminable in a closed-set identification paradigm when the spectrotemporal representation is quite sparse.

4. Discussion

There is a vast literature and long history describing various methods of distorting or recoding the speech stimulus and evaluating those effects on intelligibility. While a comprehensive review of that literature is beyond the scope of this letter, there are a few methods that are sufficiently similar to the current method as to deserve comment. As noted above, there is considerable interest in the essential aspects of speech that must be preserved in the processing provided by cochlear implants. Shannon *et al.* (1995) demonstrated that speech-envelope-modulated noise could be quite intelligible if a sufficient number of frequency bands were used. Brungart *et al.* (2005) compared three variations in speech processing and found that each method provided impressively high intelligibility, at least with a sufficient number of bands and high signal-to-noise ratios in closed-set tasks. The methods they compared were noise band or tonal carriers modulated by the speech envelopes for fixed frequency bands, and "sinewave speech" tracking formants like that used by Remez *et al.* (1981). Not surprisingly, they reported that intelligibility declined with decreasing numbers of frequency bands for all of the methods. In the current study, reducing the resolution of the speech (smaller number or longer duration of tones) degraded performance, likely due to an inadequate representation of important contrasts signaling phoneme identity. In Exp. 2, the number of contiguous bands from which the pointillistic speech was derived was varied and compared to speech processed into a subset of narrower analysis bands. Generally, the wider bands yielded better identification than the narrower bands. This finding may be related to that reported recently by Souza and Rosen (2009) who found better performance in speech recognition of sine-carrier speech for higher envelope cut-off frequencies. The processing method described here is similar in its initial stages to vocoded speech in which the carriers are pure tones modulated by narrow-band-derived envelope functions (see Arbogast *et al.*, 2002). One difference is that both frequency and amplitude are represented separately in the elements comprising the signal. However, this is not unique, either. Several investigators have demonstrated that adding slowly varying frequency modulation to

individual frequency channels previously conveying only amplitude envelope information provided speech intelligibility advantages especially in multiple talker situations (e.g., Nie *et al.*, 2005; Stickney *et al.*, 2007). This procedure is similar to the current method in that it preserves some of both the envelope and frequency information in the channels. The primary difference between that technique and the one presented here is that the within-channel variations in envelope and frequency were continuous throughout the stimulus rather than being severely quantized and replaced by equal-phase tone segments. Also similar is the procedure used by Throckmorton *et al.* (2006) in which the effect of limiting the carrier frequency to one of a small set of discrete frequencies within each channel for 2-ms time windows was investigated. Of course all digital representations of speech are quantized although the sampling rates used typically give resolution on the order of tens of microseconds. Longer time scale quantization has been used in cochlear implant coding methods and in simulated cochlear implant speech on the order of 4 ms in which the resolution in amplitude of the pure-tone carriers was varied in discrete segments or steps (Loizou *et al.*, 1999). Brungart *et al.* (2007) demonstrated that a uniform broadband noise filtered into the time-frequency regions (1/3-octave bands in 7.8 ms segments) in which the speech that it mimicked had 90% of its energy could also convey meaning. The current findings indicate that even longer discrete time segments containing both frequency and amplitude information may also result in good intelligibility under certain conditions. While it is difficult to compare performance across studies due to differences in processing as well as speech materials, Loizou *et al.* (1999) found that listeners could achieve 90% correct or better with five or more channels using open set sentence identification when the duration of each tone was 4 ms and the frequency of the sinusoidal carriers did not vary.

The nature of the speech stimulus in the current study is also similar to speech processed using a variation in the ideal binary mask for "ideal time-frequency segregation" (e.g., Brungart *et al.*, 2006, 2007; Li and Loizou, 2007, 2008; Kjems *et al.*, 2009). Analogous to the findings in Exp. 3, these authors noted that high proportions of the time-frequency bins in their approach could be removed or masked while intelligibility was retained. The similarity between the different approaches lies in the rendering of the speech stimulus into a matrix of discrete spectrotemporal units that may be analyzed or manipulated independently. However, the current method requires very little stored information to reconstruct the stimulus, only two values per matrix element. Unlike here, the ideal time-frequency algorithm usually preserves the original speech stimulus in those time-frequency units (although the Li and Loizou (2008) study also applied the binary mask technique to vocoded stimuli) and is used to separate a target source from an overlapping speech masking source using *a priori* knowledge of each prior to the mixture. It is possible, and we are currently exploring this issue, to represent two distinct sources using acoustically non-overlapping sets of tones in the current procedure. The results of Exp. 3, which demonstrated that high identification scores may be obtained in a closed-set format even when one-half or more of the elements are randomly (compared to "ideally") removed, suggest that such a multisource approach is feasible. Thus two, or perhaps even three, intelligible speech sources could be represented with mutually exclusive tonal elements. Presumably, relative level would be the primary cue in determining the most salient source although this issue awaits future investigation. Also of interest is the relative proportion of energetic and informational masking for various methods. The two techniques share many similarities and we plan future work to compare and contrast the two approaches.

Acknowledgments

References and links

Arbogast, T. L., Mason, C. R., and Kidd, G., Jr. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112**, 2086–2098.

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.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (**2006**). "Isolating the energetic component of speech-

on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Brungart, D., Iyer, N., and Simpson, B. (**2007**). "Speech perception from a crudely quantized spectrogram: A figure-ground analogy (A)," J. Acoust. Soc. Am. **121**, 3186.

Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997**). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403–2411.

Dudley, H. (**1939**). "The vocoder," Bell Lab. Rec. **18**, 122–126.

French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.

Kidd, G., Jr., Best, V., and Mason, C. R. (**2008**). "Listening to every other word: Examining the strength of linkage variables in forming streams of speech," J. Acoust. Soc. Am. **124**, 3793–3802.

Kjems, U., Boldt, J., Pedersen, M., Lunner, T., and Wang, D. (**2009**). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. **126**, 1415–1426.

Licklider, J. C. R., and Pollack, I. (**1948**). "Effects of differentiation, integration and infinite peak clipping on the intelligibility of speech," J. Acoust. Soc. Am. **20**, 42–51.

Li, N., and Loizou, P. (**2007**). "Factors influencing glimpsing of speech in noise," J. Acoust. Soc. Am. **122**, 1165–1172.

Li, N., and Loizou, P. (**2008**). "Effect of spectral resolution on the intelligibility of ideal binary masked speech," J. Acoust. Soc. Am. **123**, EL59–EL64.

Loizou, P. C., Dorman, M., and Tu, Z. (**1999**). "On the number of channels needed to understand speech," J. Acoust. Soc. Am. **106**, 2097–2103.

Miller, G. A., and Licklider, J. C. R. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**, 167–173.

Nie, K., Stickney, G., and Zeng, F.-G. (**2005**). "Encoding frequency modulation to improve cochlear implant performance in noise," IEEE Trans. Biomed. Eng. **52**, 64–73.

Poissant, S. F., Whitmal, N. A., III, and Freyman, R. L. (**2006**). "Effects of reverberation and masking on speech intelligibility in cochlear implant simulations," J. Acoust. Soc. Am. **119**, 1606–1615.

Qin, M. K., and Oxenham, A. J. (**2003**). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," J. Acoust. Soc. Am. **114**, 446–454.

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (**1981**). "Speech perception without traditional speech cues," Science **212**, 947–950.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Souza, P., and Rosen, S. (**2009**). "Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech," J. Acoust. Soc. Am. **126**, 792–805.

Stickney, G. S., Assman, P. F., Chang, J., and Zhang, F.-G. (**2007**). "Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences," J. Acoust. Soc. Am. **122**, 1069–1078.

Throckmorton, C. S., Kucukoglu, M. S., Remus, J. J., and Collins, L. M. (**2006**). "Acoustic model investigation of a multiple carrier frequency algorithm for encoding fine frequency structure: Implications for cochlear implants," Hear. Res. **218**, 30–42.

Whitmal, N. A., III, Poissant, S. F., Freyman, R. L., and Helfer, K. S. (**2007**). "Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience," J. Acoust. Soc. Am. **122**, 2376–2388.

Yang, L., and Fu, Q.-J. (**2005**). "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise," J. Acoust. Soc. Am. **117**, 1001–1004.