Jont B. Allen and Feipeng Li

Speech Perception and Cochlear Signal Processing

Speech perception is a complex process that involves multiple stages of signal processing. Once the acoustic signal reaches the human cochlear, it is decomposed into many critical bands on the basilar membrane. The cochlear nucleus then encode the temporal and frequency information in a way that is meaningful to the central auditory system.

INTRODUCTION

A major goal of speech perception research is to determine how the speech information is represented across the various stages. The research methods can be classified into three major types, i.e., the psychophysical, computational, and neurophysiological methods. The psychophysical approach [8], [9], [4], which was initiated by Harvey Fletcher and his colleagues in the 1920s, involves presenting subjects with speech stimuli and measuring their conscious responses without touching the intermediate speech decoding process within the auditory system. The computational models [10] are created for the simulation of speech perception behavior observed in psychoacoustic tests. The neurophysiological approach [7], [13] measures the detailed information of single-unit neuron response to trace the representation of speech signal through the subsequent stages of auditory processing.

After about 100 years of work, very little is know about how the ear decodes basic speech sounds. This is, in part, because it is not ethical to record in the human auditory nerve, and its not practical to do extensive speech psychophysics in nonhuman animals.

Digital Object Identifier 10.1109/MSP.2009. 932564

PERCEPTION OF SPEECH PHONEMES

The earliest studies started in the 1920s at Bell Labs, by Harvey Fletcher and his colleagues [8], [9]. Then in the 1950s, a group of speech scientists from Haskins Labs did a classic set of studies [5], [6] that indicated that speech was composed of smaller building blocks of narrow band bursts and resonances. During the 1970 and 1980s, Blumstein and Stevens [4] expanded on the Haskins work, but again, no definitive conclusions were reached, in part because synthetic speech was used, rather than real speech.

CONFUSION GROUPS

It is well known that the performance of a communication system is dependent on the symbols. The larger the "distance" between two symbols, the less likely the two will be confused, and the lower the error rate. This principle also applies to the human speech code. In a classic 1955 study [11], Miller and Nicely (MN55) collected the confusion matrices of 16 consonants /pa, ta, ka, fa, θa, sa, ʃa, ba, da, ga, va, ða, za, ʒa, ma, and na/ in white noise. The analysis of these data have shown that the consonants form natural confusion groups.

For example, Figure 1(a) shows group-average data from MN55. Each curve corresponds to the probability of reporting the labeled sound when /ta/ is presented. It is seen that /pa, ta, and ka/ form a confusion group (or perceptual group) at approximately −8 dB signal-to-noise ratio (SNR). When the SNR is decreased below −15 dB, consonant group /fa, θa, sa, and ʃa/ merges with the /pa, ta, and ka/ group, forming a super group. At very low SNRs, where no speech is audible, all the sounds asymptotically reach the chance performance of 1/16, shown by the dashed line in Figure 1(a). Thus the confusion patterns form a hierarchical structure.

A recent repeat of the MN55 study now provides similar data but, for individual utterances, and these unaveraged data shown in Figure 1(b) tell a very different story.

MORPH AND PRIME

Analysis of individual utterances reveals that as the noise level increases, certain "weak" /ta/s morph (change) into /pa/ or /ka/. For example, the /ta/ from a male talker 111 [Figure 1(b)] at 3 dB SNR, /p/ confusions overtake /t/ response. In other words, most listeners hear /pa/ rather than the target sound /ta/ as the noise in increased. This morphing effect, while at first surprising, is typical in our single consonant confusion database.

Speech tests also show that when the scores for consonants of a confusion group are similar, listeners can prime between these phones. For example, in Figure 1(b), the probabilities of /pa/ and /ta/ are equal at 3 dB SNR. At this SNR, most listeners can mentally select the consonant heard (i.e., prime), thus making a conscious choice between the two consonants. Based on our studies, it is suspected that priming occurs when events, shared by consonants of a confusion group, are at the threshold of audibility, namely when the distinguishing feature is at its masked threshold.

The fact that consonant sounds form natural groups [e.g., /p/, /t/, /k/ in Figure 1(b)] and that one sound may turn into another sound under noisy conditions clearly demonstrates that speech perception is based on discrete units. From our recent analysis of consonant confusions, the exact nature of the acoustic cues has now been discovered.

1053-5888/09/$25.00©2009IEEE

MODELING SPEECH RECEPTION

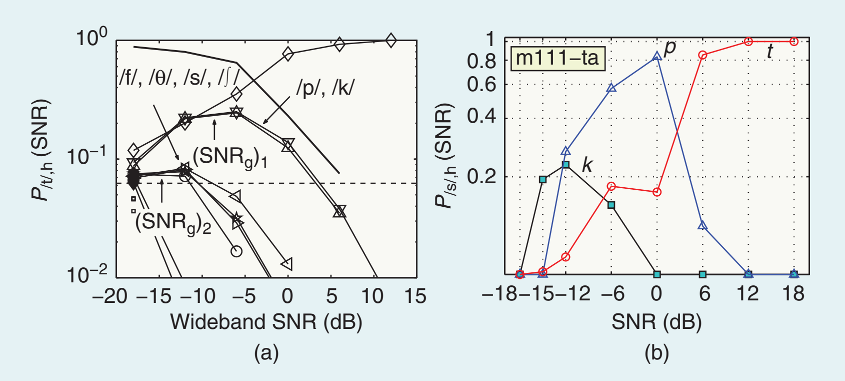
The cochlea is a nonlinear spectrum analyzer. Once a speech sound reaches the cochlea, it is represented by time-varying energy patterns across the basilar membrane (BM). A small subset of the patterns contribute to speech recognition. The purpose of event identification is to isolate this small specific feature subset.

To understand how speech sounds are represented on the BM, we have developed the AI-gram, for the visualization of the speech sounds, a what-you-see-is-what-you-hear (WYSIWYH/wisiwai/) tool, that simulates human auditory peripheral processing, The name derives from the well-known speech Articulation Index (AI), developed by Fletcher [1].

THE AI MODEL

Fletcher's AI model is an objective appraisal criterion of speech audibility. Based on the work of speech articulation over communication systems at Bell Labs, French and Steinberg developed a method for the calculation of AI [9], [8], [1]. The concept of AI is that a narrow band (i.e., a cochlear critical band) of speech frequencies carries a partial contribution to the total intelligibility, in a band-independent way. The total contribution of all bands is thus a sum of the contribution of the separate bands

$$\mathrm{AI(SNR)} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{AI}_k, \qquad (1)$$



[FIG1] **Confusion patterns (CPs) of /ta/ in white noise. (a) Average CPs of 18 /ta/s (data from MN55) reveal that /p/, /t/ and /k/ form a confusion group that are distinctive from other consonants. (b) CP of a /ta/ pronounced by a male talker m111. At 0 dB SNR most listeners report this sound as a /pa/.**

where AI_k is the specific AI for the kth articulation band, defined by

$$\mathrm{AI}_k = \min_{\mathrm{snr}}\left[\frac{1}{3}\log_{10}(1+c^2\mathrm{snr}_k^2),\, 1\right], \qquad (2)$$

and where $\mathrm{snr}_k^2 \equiv \sigma_{s+n}^2/\sigma_n^2$ is the speech to noise mean-squared (MS) ratio in the kth frequency band and $c \approx 2$ is the critical band speech-peak to noise-rms ratio [9]. Given the AI(SNR), the predicted average speech error is [1]

$$\hat{e}(\mathrm{AI}) = e_{\min}^{\mathrm{AI}}\cdot e_{\mathrm{chance}}, \qquad (3)$$

where $e_{\min}$ is the minimum error when $AI = 1$, and e_{chance} is the probability of error due to uniform guessing [1].

THE AI-GRAM

The AI-gram is the integration of a simple linear auditory model filter-bank and the Fletcher's AI model [i.e., Fletcher's SNR model of detection]. Figure 2 depicts the block diagram of AI-gram. Once the speech sound reaches the cochlea, it is decomposed into multiple auditory filter bands, each followed by an "envelope" detector. Fletcher-audibility of the narrow-band speech is predicted by the formula of specific AI (2). A time-frequency pixel of the AI-gram (a two-dimensional image) is denoted $\mathrm{AI}(t, f)$, where t and f are time and frequency. The implementation used here quantizes time to 2.5 [ms] and uses 200 frequency channels, uniformly distributed in place according to the Greenwood frequency-place map of the cochlea, with bandwidths according to the critical bandwidths of [2].

Given a speech sound, the AI-gram provides an approximate image of the effective components that are audible to the central auditory system. However, it does not label or identify those component critical for speech recognition. To find these, it is necessary to correlate the results of the speech perception experiments and the AI-grams.

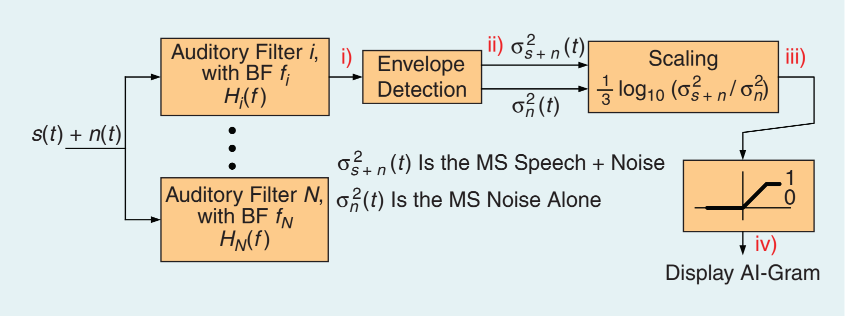
IDENTIFICATION OF CONSONANT EVENTS

A method we call the three-dimensional (3-D) approach was developed to assess the significance of audible components to speech recognition, as predicted by the AI-gram. To isolate events along time, frequency and amplitude speech sounds are truncated in time, high-/low-pass filtered, or masked with white noise, before being presented to a panel of normal hearing listeners. Once an acoustic cue critical for speech perception has been masked, the sound's recognition score is greatly reduced.

3-D APPROACH

To measure the weight of a feature to speech perception, for a particular consonant sound, the 3-D approach requires three different experiments. Each experiment had 18 talkers and between 15–25 listeners. The first experiment (TR07) determines the contribution of various time intervals by truncating the consonant into multiple segments of 5, 10, or 20 ms per frame, depending on the duration of the sound. The second experiment (HL07) divides the fullband into 12 bands of equal length along the BM, thus labeling the importance of different frequency bands by using high-pass/low-pass filtered speech as the stimuli. Once the time-frequency coordinates of the event are identified, the third experiment (MN16R) assesses the event strength by masking the speech at −18, −12, −6, 0, 6, 12, 18 dB SNR, and no noise (quiet).

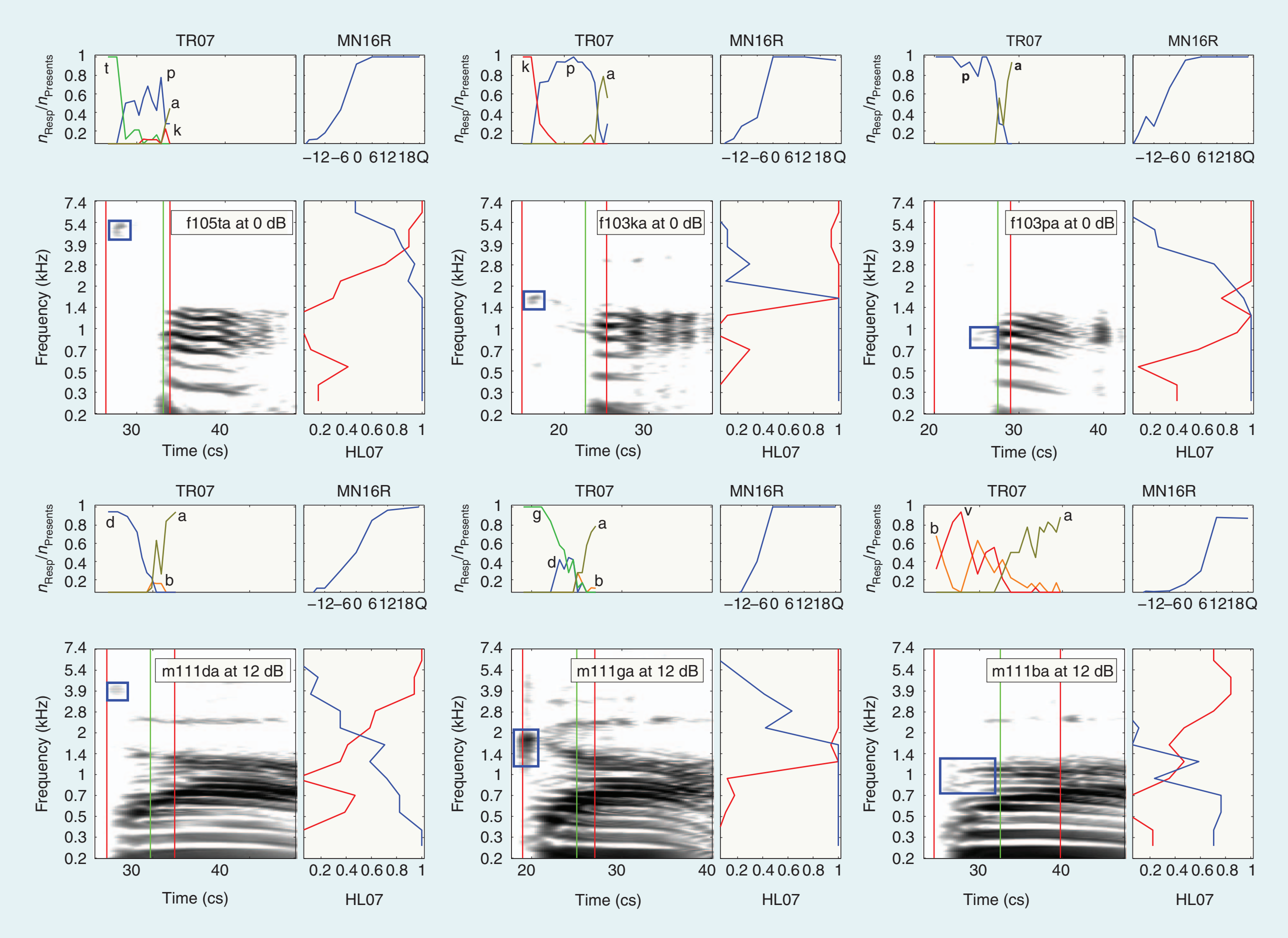
Figure 3 displays the experimental results of TR07, HL07, and MN16R, showing the probabilities of responses (the target and competing sounds), as a function of the experimental conditions, i.e., truncation time, cutoff frequency, and SNR. To facilitate the integration of the information, the results are arranged into a four-panel format, with the AI-gram depicted in the lower left panel and the recognition scores of TR07 and HL07 aligned in time (t_n in centiseconds [cs]) and frequency (along the cochlear place axis, but labeled in characteristic frequency [kHz]).


$s(t) + n(t)$
Auditory Filter i, with BF f_i $H_i(f)$
Auditory Filter N, with BF f_N $H_N(f)$
i)
Envelope Detection
ii) $\sigma^2_{s+n}(t)$
$\sigma^2_n(t)$
Scaling
iii)
$\frac{1}{3}\log_{10}(\sigma^2_{s+n}/\sigma^2_n)$
$\sigma^2_{s+n}(t)$ Is the MS Speech + Noise
$\sigma^2_n(t)$ Is the MS Noise Alone
1
0
iv)
Display AI-Gram

[FIG2] Block diagram of AI-gram.

In Figure 3, there are six sets of four panels. Each of the six sets corresponds to a specific consonant, labeled by a string that defines the gender (m, f), subject ID, consonant and SNR for the display. For example, in the upper left four panels we see the analysis of /ta/ for female talker 105 at 0 dB. Along the top are *unvoiced plosives* /t/, /k/, and /p/, while *voiced plosives* /d/, /g/, and /b/ are along the bottom. Three different talkers have been used in this analysis.

One may identify the speech event from these displays. For example, the feature that labels the sound (e.g., /t/) is indicated by the blue square in the lower-left panel of each of the six sounds (e.g., to the left of `105ta@0dB`) there is a blue box showing the burst of energy that defines the /t/ sound]. According to our 20 listeners in the TR07 experiment, when this burst is truncated, the /t/ morphs to /p/. When masking noise is added to the sound,



[FIG3] Identification of features by time and frequency bisection. Along the top are unvoiced consonants /t/, /k/, and /p/, while the corresponding voiced consonants /d/, /g/ and /b/ are along the bottom. Each of the six sounds consists of four subpanels depicted in a compact form such that the AI-gram and the three scores are aligned in time (t_n in centiseconds [cs]) and frequency. For example, for /t/ (upper left), we see four panels consisting of the time-truncation confusions (upper left), the score versus SNR (upper right), the AI-gram (lower left), and the low-pass (red) and high-pass (blue) score as a function of cutoff frequency (lower right).

such that it masks the boxed region, the percept of /t/ is lost. When the high- and low-pass filters remove the frequency of the /t/ burst, again the consonant is lost. Thus the three experiments are in agreement, and they uniquely isolate the location of the event responsible for /t/.

This nicely generalizes to the other plosive consonants shown (i.e., /t/, /k/, /p/, /d/, /g/, and /b/). From such data we see that /t/ is labeled by a 4 kHz burst of energy ≈50 ms before the vowel, whereas /k/ is defined as a 1.4–2 kHz burst, also ≈50 ms before the vowel. A burst of energy leading the vowel at 0.7–1 kHz defines the /p/. The three voiced sounds /d/, /g/, and /b/ have similar frequencies but are presented at the same time as the vowel onset.

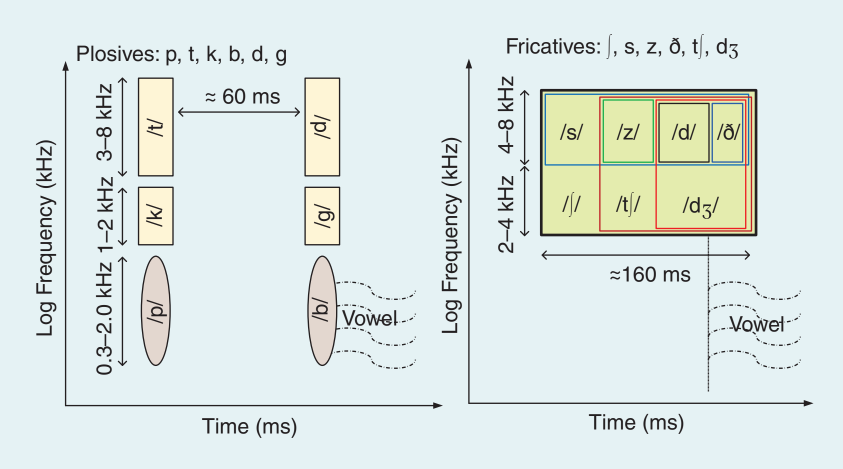
The two high-frequency sounds (top and bottom left of Figure 3) are /t/ and /d/, each produced with the tongue tip on the roof of the mouth and slightly behind the teeth. The two midfrequency sounds, /k/ and /g/ are produced with the back of the tongue, labeled in the frequency domain as bursts between 1.4–2 kHz, for the examples shown. Finally low-frequency /p/ and /b/ are produced with the release of the lips. These two sounds produce a very low-frequency burst between 0.7 and 1 kHz. The exact relationship between the place of the burst feature and the burst frequency needs further explanation.

We have analyzed all 128 sounds in our consonant ID database and similar results have been found. Thus, we are confident that these tags of energy label the identity of these consonants. The distributions of the burst frequencies, durations, and delays to voicing need further study, especially for the case of vowels other than /a/, used here.

OVERVIEW OF CONSONANT EVENTS

Figure 4 provides a schematic drawing of the events of initial consonants preceding vowel /a/. The stop consonants are characterized by the center frequency of the burst, caused by the sudden release of pressure in the oral cavity. Besides the voice bar, the voiced and unvoiced stops differ mainly in the duration of the transition. The fricatives are characterized by an onset of wideband noise created by the turbulent airflow through lips and teeth. Duration and frequency range are the two critical parameters for fricatives. A voiced fricative usually has a considerably shorter duration than its unvoiced counterpart. The events of the consonants are consistent across the different talkers, despite that the parameters, such as timing, frequency, and strength may change, to a certain degree, within the given range.

To further verify these results, we have developed a method to modify the speech sounds using the short-time Fourier transform (STFT) [3]. For the stop consonants, it is shown that /pa/, /ta/, and /ka/ can be converted into each other simply by modifying the burst amplitude. Removing the /ka/ burst in the midfrequency turns a /ka/ into a /ta/ or /pa/; boosting the onset in the high frequency turns the weak /t/ sound into a noise-robust /ta/; the same sound turns into a well-articulated and natural /pa/ if both the onsets in the high frequency and midfrequency are removed. Similar conversions can be made to the voiced stop consonants, /ba/, /da/, and /ga/. For the fricatives, we demonstrate the conversion of /ʃa/ → /sa/ → /za/ → /ða/ simply by cutting the duration or bandwidth of the frication region. Examples of these modifications may be found at http://hear.ai.uiuc.edu/wiki/Files/VideoDemos.



[FIG4] Structure of the plosives and the fricatives, in terms of time-frequency allocation. Mapping these regions into events requires extensive perceptual experiments. But once the sounds have been evaluated, it is possible to prove where the key noise-robust events live in perceptual space.

COCHLEAR SIGNAL PROCESSING

The cochlea plays a vital role in speech perception. Once the cochlea is damaged, the ability to process speech in noise is seriously degraded. The main functions of cochlea are to separate the input acoustic signal into overlapping frequency bands and to compress the large acoustic intensity range into the much smaller mechanical and electrical dynamic range of the inner hair cell [2]. The auditory neurons then convert the signal into neural spikes and send them to the auditory system. This raises the basic question of information processing by the ear. The eye plays a similar role as a peripheral organ. It breaks the light image into rod- and cone-sized pixels, as it compresses the dynamic range of the visual signal. Based on the intensity just-noticeable difference (JND), the corresponding visual dynamic range is about 9–10 orders of magnitude of intensity, while the ear has about 11–12 [2]. Neurons are low-bandwidth channels. The stimulus has a relatively high information rate. The eye and the ear must cope with the bandwidth problem by reducing the stimulus to a large number of low-bandwidth signals. It is then the job of the cortex to piece these pixelized signals back together, to reconstruct the world as we see and hear it.

SENSORINEURAL HEARING LOSS

Most sensorineural hearing loss can be attributed to the malfunction of cochlear outer hair cells (OHCs) and inner hair cells (IHCs). Damage to OHCs reduces the vibration of the cell's cilia at the stimulus frequency, resulting in an elevated detection threshold. Damage to the IHCs reduces the efficiency of mechanical-to-electrical transduction, also resulting in an elevated detection threshold. The audiometry configuration is not a good indicator of the physiological nature of the hearing loss [12], specifically, subjects with OHC and IHC loss may show the same amount of shifting in hearing threshold, yet the influence of the two types of hearing loss on speech perception can be very different.

The loss of IHCs has a serious impact on speech perception, as supported by the results of an elderly subject (AS) with moderate hearing loss, who volunteered in our pilot study of hearing-impaired speech perception. Due to a cochlear dead region (an extreme case of IHC loss [12]) from 2–3.5 kHz, where the perceptual cues for /ka/ and /ga/ are located, AS cannot hear these two sounds with her left ear. In contrast, her right ear can hear /ka/ and /ga/ (with low accuracy), despite the fact that the two ears have an almost identical hearing threshold. A consonant confusion analysis shows that more than 80% of the /ka/s are misinterpreted as /ta/, while about 60% of the /ga/s are reported as /da/.

It is well known that noise damage of "nerve cells" (i.e., OHCs) leads to *loudness recruitment*, the most common form of neurosensory hearing loss, characterized as the reduction in dynamic range. To successfully design hearing aids that deal with the problem of recruitment, we need models to improve our understanding of how the cochlea achieves its dynamic range. Given the observations shown here as speech events, we need to extend our primitive understanding of wide-dynamic range compression into the time domain.

It is also conjectured that speech onsets will be enhanced by OHC processing, due to the overshoot observed in the auditory nerve timing. In the hearing impaired ear, such enhancements would be gone, therefore this extra "kick" of response would not be available in those ears.

SUMMARY

Speech sounds are encoded by time-varying spectral patterns called acoustic cues. The processing and detection of these acoustic cues lead to events defined as the psychological correlates of the acoustic cues. Due to the similarity between the acoustic cues, speech sounds form natural confusion groups. When the feature of the sound within a group is masked by noise, one event can turn into another. A systematic psychoacoustic "3-D method" has been developed to explore the perceptual cues of stop consonants from naturally produced speech sounds. For each sound, our 3-D method measures the contribution of each subcomponent by time-truncating, high-pass/low-pass filtering, and masking with noise. The AI-gram, a visualization tool that simulates the auditory peripheral processing, is used to predict the audible components of the speech sound. The results are that the plosive consonants are defined by a short duration bursts characterized by their center frequency, as well as the delay to the onset of voicing. Fricatives are characterized by the duration and bandwidth of a noise-like feature. Pilot studies of hearing-impaired (HI) speech perception indicate that cochlear dead regions have a considerable impact on consonant identification. An HI listener may have problems understanding speech simply because he/she cannot hear certain sounds, since the events are missing due to either the hearing loss, or the masking effect introduced by the noise.

ACKNOWLEDGMENTS

We gratefully acknowledge Phonak and the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign for their generous support of this research. A small portion of this research has been supported by NIH Grant RDC009277A.

AUTHORS

Jont B. Allen (jontalle@illinois.edu) received a B.S. degree in electrical engineering from the University of Illinois, Urbana-Champaign in 1966, and the M.S. and Ph.D. degrees from the University of Pennsylvania in 1968 and 1970, respectively. From 1974 to 1997, he was in the Acoustics Research Department in Bell Laboratories, Murray Hill, New Jersey. He was with the research division of the newly created AT&T Labs from 1997 to 2002. Since 2003, he has been with the University of Illinois at Urbana-Champaign.

Feipeng Li (fli2@illinois.edu) is a Ph.D. candidate in the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign. He received his B.S. and M.S. degrees, both in electrical engineering from Wuhan University, China, in 1996 and 1999, respectively. After graduation, he joined the National Remote Sensing Lab at Wuhan University, where he was a research scientist.

REFERENCES

[1] J. B. Allen, "How do humans process and recognize speech?" *IEEE Trans. Speech Audio.*, vol. 2, no. 4, pp. 567–577, 1994.

[2] J. B. Allen, "Nonlinear Cochlear Signal Processing and Masking in speech perception," in *Springer Handbook on Speech Processing and Speech Communication*, J. Benesty and M. Sondhi, Eds. Germany: Springer-Verlag, 2008, ch. 3, pp. 27–60.

[3] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[4] S. E. Blumstein and K. N. Stevens, "Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants," *J. Acoust. Soc. Amer.*, vol. 66, no. 4, pp. 1001–1017, 1979.

[5] F. Cooper, P. Delattre, A. Liberman, J. Borst, and L. Gerstman, "Some experiments on the perception of synthetic speech sounds," *J. Acoust. Soc. Amer.*, vol. 24, no. 6, pp. 579–606, 1952.

[6] P. Delattre, A. Liberman, and F. Cooper, "Acoustic loci and translational cues for consonants," *J. Acoust. Soc. Amer.*, vol. 27, no. 4, pp. 769–773, 1955.

[7] B. Delgutte, "Auditory neural processing of speech," in *The Handbook of Phonetic Sciences*, W. J. Hardcastle and J. Laver, Eds. Oxford, U.K.: Wiley-Blackwell, 2002, pp. 507–538.

[8] H. Fletcher and R. Galt, "Perception of speech and its relation to telephony," *J. Acoust. Soc. Amer.*, vol. 22, no. 2, pp. 89–151, 1950.

[9] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, no. 1, pp. 90–119, 1947.

[10] J. L. McClelland and J. L. Elman, "The trace model of speech perception," *Cognitive Psychol.*, vol. 18, pp. 1–86, 1986.

[11] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Amer.*, vol. 27, no. 2, pp. 338–352, 1955.

[12] B. C. J. Moore, "Dead regions in the cochlea: Conceptual foundations, diagnosis, and clinical applications," *Ear Hearing*, vol. 25, no. 2, pp. 98–116, 2004.

[13] S. A. Shamma, "Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve," *J. Acoust. Soc. Amer.*, vol. 78, no. 5, pp. 1612–1621, 1985.