Manipulation of Consonants in Natural Speech

Feipeng Li and Jont B. Allen, *Life Fellow, IEEE*

***Abstract*—Natural speech often contains *conflicting cues* that are characteristic of confusable sounds. For example, the /k/, defined by a mid-frequency burst within 1–2 kHz, may also contain a high-frequency burst above 4 kHz indicative of /ta/, or vice versa. Conflicting cues can cause people to confuse the two sounds in a noisy environment. An efficient way of reducing confusion and improving speech intelligibility in noise is to modify these speech cues. This paper describes a method to manipulate consonant sounds in natural speech, based on our *a priori* knowledge of perceptual cues of consonants. We demonstrate that: 1) the percept of consonants in natural speech can be controlled through the manipulation of perceptual cues; 2) speech sounds can be made much more robust to noise by removing the conflicting cue and enhancing the target cue.**

***Index Terms*—Conflicting cue, perceptual cue, speech processing.**

I. Introduction

After a half century of study, many speech processing techniques such as synthesis, noise reduction, and automatic speech recognition (ASR), have reached a plateau in performance. For example, the performance of the state-of-the-art ASR systems is still far below that of human speech recognition (HSR) [18]. A major problem is that it is fragile under noisy conditions. The best phone classification accuracy in ASR systems varies from 82% in quiet [35] to chance performance at 0 dB signal-to-noise ratio (SNR). For HSR, the average phone classification accuracy in quiet is near 98%–98.5% (1.5%–2% error) [3], [4], while the SNR required for chance performance is below −20-dB SNR [50]. For many sounds, the phone classification performance in humans is unchanged from quiet to 0-dB SNR [51] in white noise. In the past, ASR research has benefited significantly from HSR research. For instance, the use of delta Mel-frequency cepstral coefficients (MFCCs) as the feature vector was rationalized by the perceptual study on time-truncated syllables [30]. It is now widely accepted that bio-inspired speech processing schemes have the potential to lead to better solutions for noise-robust speech recognition [53], [18], [34] and other applications.

Manuscript received November 04, 2009; revised February 18, 2010; accepted April 23, 2010. Date of publication July 12, 2010; date of current version October 29, 2010. This work was supported in part by the National Institute of Health under Grant RDC009277A, awarded 07/31/2008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Malcolm Slaney.

F. Li is with the Biomedical Engineering Department, Johns Hopkins University, Baltimore, MD 21205 USA (e-mail: fli12@jhmi.edu).

J. B. Allen is with the Electrical and Computer Engineering Department, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: jontalle@illinois.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2050731

Perhaps ASR performance will improve if we can answer the fundamental question of HSR: *How is the speech coded in the auditory system?* In order to determine the basic speech spectral patterns, in 1952, Cooper and Liberman and their colleagues built a machine called the *pattern playback* that generated artificial speech from spectrograms, and then went on to conduct a classic series of psychoacoustic studies on the perception of synthetic stop consonants [14], [16]. Later, the method of speech synthesis was widely used in the search for acoustic correlates for stops [10], [37], fricatives [36], [33], nasals [46], [43], [54], and distinctive or articulatory features [11], [12], [61]. A major drawback of this method is that to synthesize speech, it requires the experimenter to have know *a priori* knowledge about the speech cues to be identified. In fact, the speech stimuli generated by the speech synthesizers, such as *pattern playback*, are generally of low quality, even barely intelligible, because the assumptions about the features are either incomplete or inaccurate. To identify cues in natural speech, it is necessary to have a direct way of measuring them. Of course, this has been the difficult challenge. [10], [11], [18], [19], [37]

To understand how speech information is represented in the human auditory system, a number of researchers have studied the recordings of single auditory neurons in animals in response to speech stimuli [19], [58]. Since it has been unethical to record in the human auditory nerve, and it is difficult to do extensive speech psychophysics in nonhuman animals, those aforementioned neurophysiological studies were unable to be correlated with human psychophysical data. We have dealt with this problem by creating a computational model of speech reception, called the AI-gram [45], [55], by integrating Fletcher's Articulation Index (AI) model of speech intelligibility [29], [27], [39], [2] and a simple linear auditory model filter-bank (i.e., Fletcher's critical-band SNR model of detection [3]). Given a speech sound in noise, the AI-gram provides an initial estimate of audibility of various time–frequency components in the central auditory system. However, just because a component is audible does not mean it is information bearing. We have found that large portions of audible speech are *not* information-bearing. When these portions are removed, the quality or timbre of the speech changes, but not the conveyed meaning. To address this issue, a systematic psychoacoustic method, denoted the *three-dimensional deep search (3DDS)*, has been developed to identify true information-bearing events [41], [7]. The core idea behind 3DDS is to systematically remove various parts of a speech sound and then to assess the importance of the removed component from the change in the recognition score. In order to measure the distribution of speech information along the time, frequency, and amplitude dimensions, three different and independent psychoacoustic experiments are per-

formed on each speech token. Each experiment consists of one of the following independent methods: 1) speech sounds are truncated in time; 2) high-/low-pass filtered in frequency; 3) masked with white noise. The modified sound stimulus is presented to a battery of about 20 normal hearing listeners, with trials randomized across utterances and conditions [7], [41], [42]. Once an event is removed through time-truncating, filtering and masking, the recognition score of human listeners drops abruptly [55], [41]. As a quantitative way of measuring speech cues, the 3DDS has at least two major advantages over the conventional methods [14], [16], [11]. First, 3DDS uses natural speech; thus, the method makes no tacit assumption about the relevant cues. Second, it harnesses the large variability of natural speech. More than 18 talkers and listeners are employed in each of the three experimental procedures to carefully sample talker-listener space. The information from the three experiments was then combined to create a single estimate of each event. This approach has proven successful when applied to initial consonant–vowel (CV) sounds for both plosives [41] and fricatives [7], [48].

We have discovered that naturally produced consonants often contain conflicting cues, which are the sources of consonant confusions [42], once the dominant cues that define the target sounds are masked. Through the manipulation of the dominant/conflicting cues, usually just a small time–frequency region in the AI-gram, we can morph one phone into another, demonstrating that speech perception is critically dependent on these perceptual cues. Moreover, the robustness (intelligibility) of consonants in noise is determined by the relative intensity of the perceptual and conflicting cues [41].

These observations of HSR impose important implications for both automatic speech recognition and speech enhancement; first, a perceptual-cue-based processing scheme might provide improved robustness or intelligibility of consonants in noise. As we mentioned earlier, current ASR systems fail with even small amounts of masking noise that have little or no impact on HSR. Many researchers believe that it is because the front-end does not resolve the features that are resilient to ambient noise. Second, the existence of conflicting cues in natural speech further complicates the training of ASR systems. Over the past years, various noise-reduction techniques have been proposed to increase the SNR [22], [40], but none of these methods have been shown to be effective in improving speech intelligibility [8], [9]. A more effective way might be to work directly with the perceptual cues.

Here we present a method of manipulating consonant sounds in natural speech, based on our *a priori* knowledge of perceptual cues of consonants [41], [48], and demonstrate its potential use for noise-robust speech recognition. The paper is organized as follows. Section II gives an overview of the perceptual cues for consonant sounds. Section III shows how the percept of naturally produced consonants may be manipulated through the operations on acoustic cues. Section IV tests the idea of noise-robust consonant recognition with a psychoacoustic experiment, and in Section V we summarize our findings and discuss the limitations of our current method.

II. Perceptual Cues of Consonant Sounds

In natural speech, due to the physical constraints on the articulators (mouth, tongue, lips, etc.), it is widely accepted that their "ideal" position are often compromised due to neighboring sounds (e.g., a V on a C). As a consequence, speech cues of successive $\{C, V\}$ units frequently interact, an effect called *coarticulation* [28]. Since coarticulation does not extend beyond neighboring syllables, it is allowable to separate continuous speech into syllable segments, such as CV or CVC [49].

Using the 3DDS method, we have identified the perceptual cues of initial consonants preceding vowel /a/, /i/, and /u/ [42], [48], [7].

A. Overview of Consonant Cues

Fig. 1 depicts the AI-grams of 16 consonants preceding vowel /a/, with the dominant perceptual cues highlighted by the rectangular frames. The stop consonants /p, t, k, b, d, g/ are characterized by a compact burst of short duration (less than 15 ms) caused by the sudden release of pressure in the oral cavity. Within the same group, the stop consonants distinguish themselves by the center frequency of the burst, specifically /ta/ and /da/ are labeled by a high-frequency burst above 4 kHz; /ka/ and /ga/ are defined by a mid-frequency burst from 1.4–2 kHz, whereas /pa/ and /ba/ are represented by a soft wide-band click, which often degenerates into a low-frequency burst from 0.7–1 kHz due to the masking effect of surrounding noise. The voiced and unvoiced stops differ mainly in the duration of the gap between the burst and the start of sonorance. The fricatives /f, s, ʃ, ʧ, v, z, ʒ, ʤ/ are characterized by a salient noise-like cue caused by the turbulent air flow through constrictions in the lips, teeth and palate. Duration and bandwidth are two key parameters for the discrimination of these sounds. Specifically, the /fa/ cue is within 1–2.8 kHz and lasts for about 80 ms; the /sa/ cue falls within 4–8 kHz and lasts for about 160 ms; /ʃa/ is also labeled by a cue of long duration, but it has a lower frequency (2–4 kHz); the /ʧa/ cue ranges from 2–8 kHz and lasts for more than 100 ms. These results are summarized from [41], [48]. The voiced fricatives have similar patterns of perceptual cues, except that the durations are considerably shorter than their unvoiced counterparts. The two nasals /m/ and /n/ share a common feature of nasal murmur at low frequency and differ from each other in their mid/low-frequency timing and F2 onset (below 2.4 kHz).

These invariant consonant cues have been found to be systematic across talkers. Similar 3DDS data for two other vowels /i/ and /u/ is currently being analyzed. In running speech, the acoustic cues are expected to change depending on the preceding and following vowels [14].

B. Conflicting Cues

Due to the physical limitations of the human speech articulators, it is difficult to produce "ideal" speech sounds, such as those generated by a speech synthesizer. We have found that many natural CV sounds contain conflicting cues indicative of competing sounds. Our analysis of the Linguistic Data Consortium (LDC) LDC2005S22 "Articulation Index Corpus" (University of Pennsylvania) indicates that most stop consonants /pa,

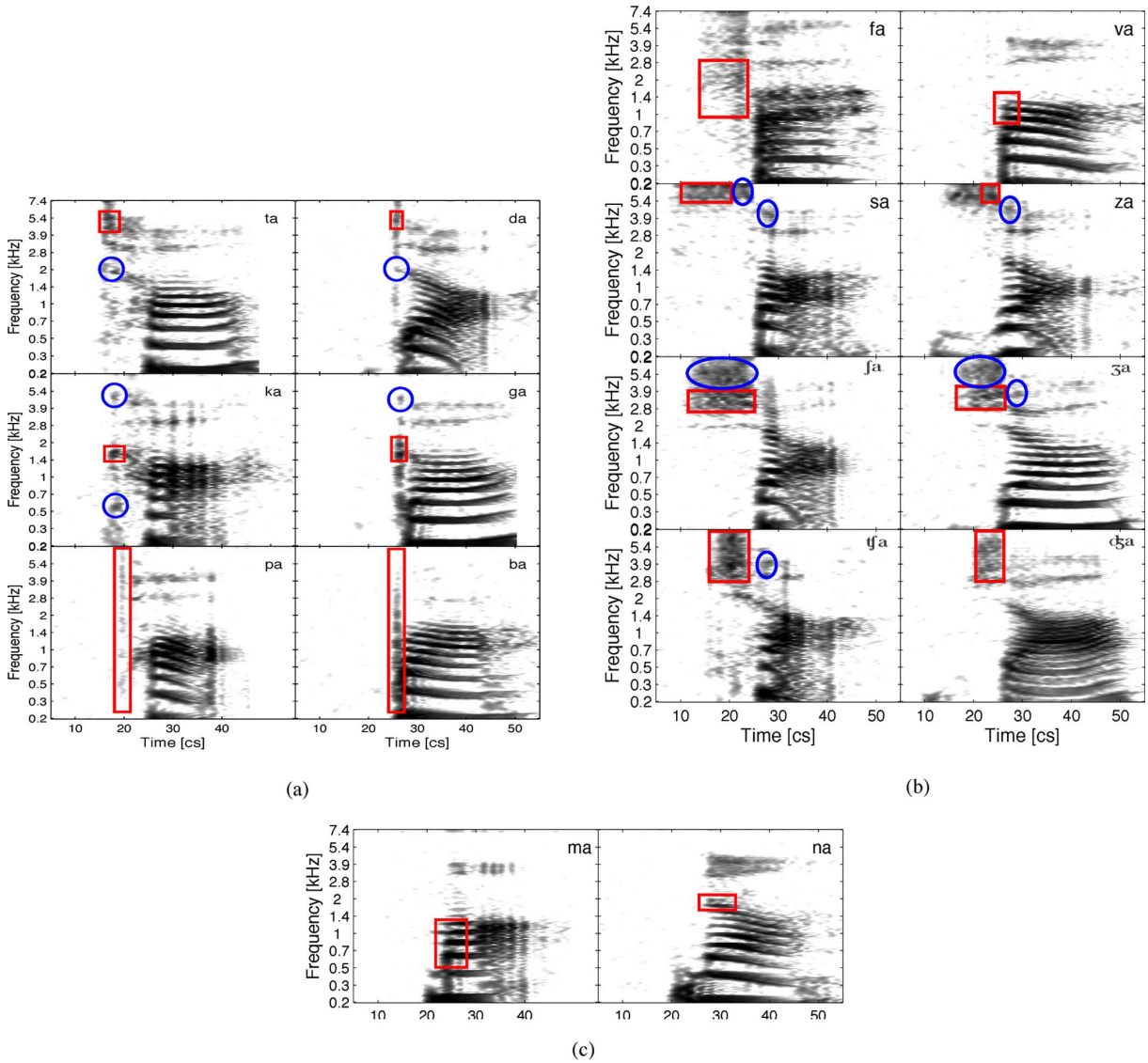

Fig. 1. AI-grams for the 16 Miller–Nicely consonants at 12-dB SNR in white noise. (a) Stops. (b) Fricatives. (c) Nasals. All sounds are pronounced by female talker f103 except for /fa/, which is produced by talker f101. A rectangular frame highlights the perceptual cue that distinguishes each sound from its competing sounds, as determined by the 3DDS procedure [42], [7]. The conflicting cues are labeled by ellipses. These AI-grams form a baseline starting point for speech modifications of the boxed regions. (1 cs $= 0.01$ s). (a) Stops: /ta, ka, pa, da, ga, ba/. (b) Fricatives: /fa, sa, ʃa, ʧa, va, za, ʒa,ʤa/ (c) Nasals: /ma, na/.

ta, ka, ba, da, ga/ contain combinations of consonant cues that may lead to confusions in speech perception under adverse circumstances. As an example, /ka/ from talker f103 is shown in Fig. 1(a). The talker (f103) intends to produce a /ka/ phone, and the listeners report hearing /ka/ 100% of the time at 0 dB in both white noise (WN) and speech weighted noise (SWN) and a notable 98% of the time at −10-dB SNR in SWN. Yet, the produced speech contains both a high-frequency burst around 5 kHz (indicative of a /ta/ production) and a low-frequency burst spanning 0.4–0.7 kHz (indicative of a /pa/ production), as indicated by the circles in the figure. When these two conflicting cues are digitally removed, one hears no difference between the modified sound and the original sound. In this example, the listeners report a robust /ka/ because the mid-frequency /ka/ burst (highlighted by a rectangular box) perceptually "overpowers" the conflicting cues. Exactly how this happens is not understood, but it is a result of cochlear and neural processing of the auditory nerve signal. This effect is shown for /ga/ in Fig. 1(a). In addition to the typical /ga/ burst in the mid-frequency (highlighted by a rectangular box), this speech sample also contains a high-frequency burst above 4 kHz (labeled by a circle), which could result in a /ga/→/da/ confusion, if the /g/ burst is masked or removed.

Conflicting cues also exist in fricative consonants. As seen in Fig. 1(b), the fricative time section of /ʃa/ also contains a /sa/ cue above 4 kHz (labeled by an ellipse). Similarly, within the fricative time section of /sa/ we also see the perceptual cue for /za/. Apart from these examples, /sa, ʃa, ʧa, za, ʒa/ all contain a high-frequency burst above the head of the F3 transition (labeled by ellipses); this cue, if presented alone, could lead to the perception of /ða/. As before, if the conflicting cue is removed, the sound is literally indistinguishable from the unmodified speech.

Because of the existence of conflicting cues, the percept of a sound predictably changes if the dominant cue is masked. This effect is further described in Section III, where we discuss the manipulation of consonants in natural speech.

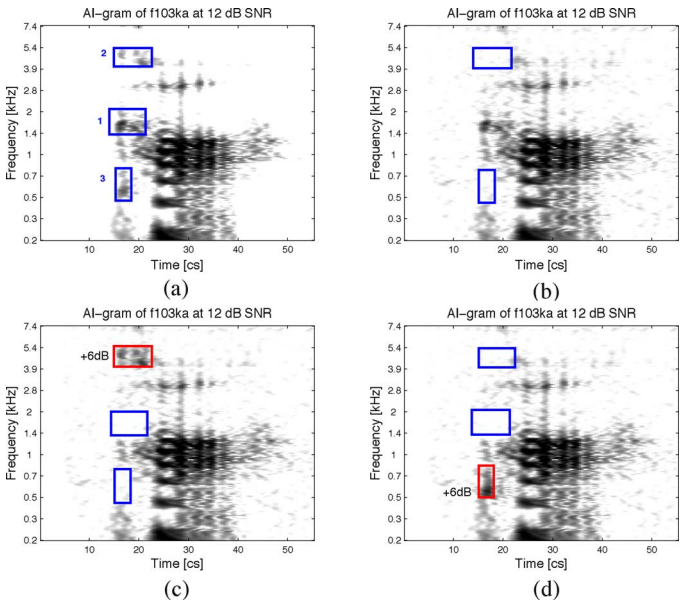

Fig. 2. Three-way manipulation of unvoiced stop consonant /ka/. (a) Original /ka/ from talker f103 at 12-dB SNR. (b) When the two conflicting cues (blocks 2 and 3) are removed, one hears no difference. (c) When block 1, containing the /k/ cue, is removed and the /t/ cue (block 2) is enhanced by 6 dB, a /t/ is robustly reported. (d) When both the /k/ and /t/ cues are removed (blocks 1 and 2), /pa/ is robustly reported. [Example: "ka→ka→ta→pa"]. (1 cs = 0.01 s).

III. Manipulation of Speech Cues

Speech perception is a complex multilevel process where the integration of events is governed by high-level language such as lexical, morphological, syntactic, and semantic context. To manipulate phones in natural speech, it is convenient to start from nonsense syllables, so that the high-level constraints on speech perception are maximally controlled [5]. We first examine the manipulation of initial consonants as they occur in isolated nonsense CV syllables. We then show that speech cues may be modified in isolated meaningful syllables (words) and sentences. The examples discussed in this report can be found at http://hear.ai.uiuc.edu/wiki/Files/VideoDemos. For example, the sample "ka→ka→ta→pa" from Fig. 2 is listed as "ka2ka2ta2pa" on the website.

Our speech modification procedure begins by analyzing the speech sounds using the short-time Fourier transform (STFT). The boxed regions of Fig. 1 are modified, and the modified speech is then returned to the time domain via an overlap-add synthesis [1].

A. Speech Analysis and Synthesis

Let $s[n]$ denote the speech signal at sample times n. For analysis, the original signal $s[n]$ is divided into N point overlapping frames $s[m,n] \equiv w[n]s[mR-n]$ of 20-ms duration with a step size $R \equiv N/4$ samples of 5 ms. A Kaiser window $w[n]$ having -91 dB attenuation (i.e., first side lobe is 91 dB smaller than the main lobe) is used. Note that the speech is time-reversed and shifted across the fixed window prior to being Fourier transformed

$$X[m,k] = \sum_{n=0}^{N-1} s[m,n]e^{-j2\pi kn/N}. \tag{1}$$

The resulting STFT coefficients $X[m,k]$ is a two-dimensional complex signal matrix, indexed in time m and frequency k.

The region of a speech cue is modified by multiplying $X[m,k]$ with a two-dimensional mask $M[m,k]$ that specifies the gain g within the feature area. Specifically, $g=0$ is feature removal, a gain $0<g<1$ corresponds to a feature attenuation, while a gain $g>1$ is feature enhancement, resulting in the modified speech spectrum

$$Y[m,k] = X[m,k]\cdot M[m,k]. \tag{2}$$

The gain may be expressed in dB as $G = 20*\log_{10}(g)$ dB. Following modifications, the single frame signal can be recovered by applying an inverse Fourier transform

$$y[m,n] = \frac{1}{N}\sum_{k=0}^{N-1} Y[m,k]e^{j2\pi kn/N} \tag{3}$$

followed by the overlap add (OLA) synthesis, resulting in the modified speech signal $y[n]$

$$y[n] = \sum_{m=-M_0}^{0} y[mR,n] \tag{4}$$

over all past samples [1].

To improve the accuracy of modification, the windowed speech is zero-padded before performing the Fourier transform.

B. Nonsense Syllable

1) Plosives: To demonstrate that the unvoiced stop consonants /pa/, /ka/, and /ta/ can be converted from one to the other (because of the conflicting cues), we select a /ka/ from talker f103, the same example discussed in Section II-B. Using the signal processing method described in Section III-A, we modify the speech by varying the relative levels of three speech cues (highlighted by the three blocks in Fig. 2). When the mid-frequency /ka/ burst in block 1 is removed [Fig. 2(a)], the percept of /ka/ is dramatically changed and listeners report either /pa/ or /ta/. This ambiguous situation leads to *priming*, which is defined as the auditory illusion where prior expectation of the perceived sound affects the sound reported. In other words, for this illusion a listener can consciously switch between two or more choices thus predecide the consonant being heard. When both short bursts for /ka/ and /ta/ (blocks 1, 2) are removed, the sound is robustly perceived as /pa/. Boosting the low-frequency burst within 0.5 and 0.7 kHz (block 3) strengthens the initial aspiration and turns the sound into a clearly articulated /pa/ [Fig. 2(d)] (which may not be primed).

An interesting question about this example is: why do people hear /ka/ rather than /ta/ and /pa/? We conjecture that it is be-

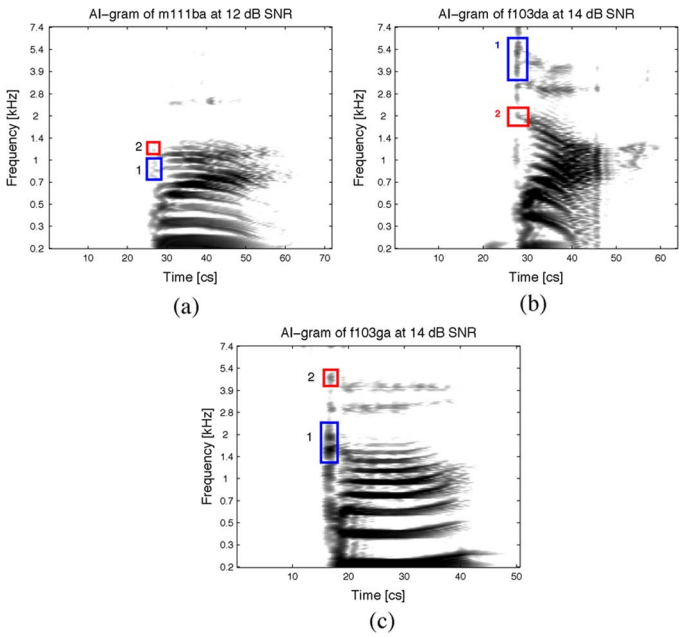

Fig. 3. Manipulation of voiced stop consonants /ba, da, ga/. (a) /ba/ from talker m111 morphs into /ga/ when the /ba/ cue in block 1 is replaced by a /ga/ cue in block 2. [Example: `ba2ga`] (b) /da/ from talker f103 is heard as a natural /ga/, after removing the high-frequency burst (block 1)and boosting the mid-frequency burst (block 1) by a factor of 5 (14 dB). [Example: `da2ga`] (c) Removal of the mid-frequency burst (block 1) causes the original sound /ga/ from talker f103 to morph into a /da/. Boosting the high-frequency burst (block 2) makes the sound a clear /da/. [Example: `ga2da`]. (1 cs = 0.01 s). (a) /ba/ → /ga/ (b) /da/→ /ga/ (c) /ga/→ /da/.

cause of the 1.4 kHz burst, which triggers the /ka/ report, renders the /ta/ and /pa/ bursts inaudible, possibly due to the upward-spread of masking or some neural signal processing mechanism.

An important implication of this example (Fig. 2) is that the F2 transition for /ka/ seems unnecessary for the discrimination of unvoiced stop consonants, contradictory to a widely accepted argument that the F2 transition is critical for the recognition of stop consonants [16], [12].

The group of voiced stop consonants /ba, da, ga/ and the unvoiced stop consonants /pa, ta, ga/ have similar feature patterns, with the main difference being the delay between the *voicing* (i.e., the burst release and the start of the sonorant portion of the speech sound). We shall next show how the voiced stops /ba, da, ga/ can be modified, again through speech cue manipulations.

Fig. 3(a) depicts the AI-gram of /ba/ from talker m111 at 12-dB SNR with white noise, which is perceived robustly by the listeners as a /ba/ above 12-dB SNR. After removing the perceptual cue for /ba/ (block 1) and boosting the mid-frequency burst (block 2) by a factor of 4 (12 dB), the speech sample is transformed into a noise-robust /ga/. Fig. 3(b) shows the AI-gram of /da/ from talker f103 at 14-dB SNR with white noise, which contains a typical high-frequency /da/ burst (block 1) and a conflicting mid-frequency /ga/ burst (block 2). Just as in Fig. 2 where /ka/ is converted to /ta/ or /pa/, the /da/ sound may be converted into a /ga/ by removing the high-frequency burst (block 1) and scaling up the lower frequency burst (block 2) to create a fully audible mid-frequency burst.

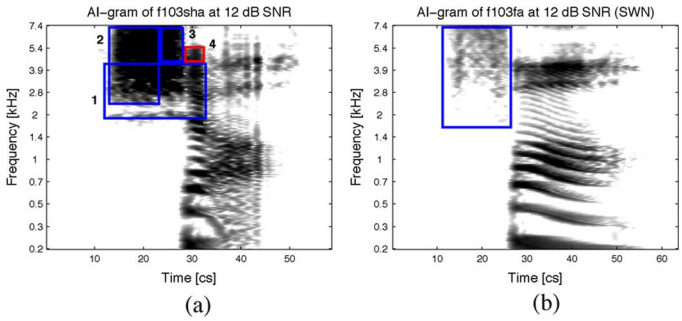

Fig. 4. Manipulation of fricatives /ʃa, fa/. (a) the original sound /ʃa/ from talker f103 is converted into a /sa/ when the bandwidth of the noise-like cue is cut from 2–4 kHz (removing block 1); it turns into a /ʧa/ when the duration is shortened from its natural duration of 15 cs (from 13–28 cs) down to 6 cs (from 22–28 cs) (removing block 2), combining the two processes (removing block 1 and 2) turns the sound into a /za/; Finally when all three blocks are taken out, the sound is heard as a /ða/, and boosting the high-frequency residual (block 4) makes the /ða/ clearer. [Example: `Sa2cha2sa2za2Da`] (b) the original sound /fa/ from talker f103 turns into a /ba/ when the whole fricative cue (highlighted by the blue box) is deleted. [Example: `fa2ba`. (1 cs = 0.01 s). (a) /ʃa/ → /sa/ → /ʧa/ → /za/ → /ða/ (b) /fa/ → /ba/.

The reverse conversion (from /ga/ to /da/) is illustrated in Fig. 3(c). After removing the mid-frequency /ga/ cue (block 1), the listeners robustly report /da/. This final modification, for some SNR conditions (when the mid-frequency boost is removed and there is insufficient high-frequency residual energy for the labeling of a /da/), requires a 12-dB boost of the 4-kHz region to robustly convert the sound to /da/.

2) Fricatives: The fricatives are characterized by a wideband noise-like cue with varied duration and bandwidth [48]. Truncating the speech cues in bandwidth and duration, we can also morph the fricatives from one into the other. Starting with /ʃa/ from talker f103 [Fig. 4(a)], the original sound is heard by all listeners as a solid /ʃa/. In the figure, the perceptual cue ranges from 13–28 cs in time and about 2–8 kHz in frequency. Cutting the bandwidth in half (remove block 1) morphs the sound into a robust /sa/. Shrinking the duration by 2/3 (remove block 2) transforms the sound into a /ʧa/. Combining both processing (remove block 1 and 2) causes most listeners to report /za/. Removing the whole noise patch (remove block 1, 2 and 3) results in /ða/, which can be made robust by amplifying the residual high-frequency burst (highlighted in block 4). In each case, the modified speech is naturally sounding.

Consonants /fa/ and /va/ are highly confused with /ba/ when the fricative sections of the two sounds are masked. Fig. 4(b) shows an example of a /fa/→/ba/ conversion. The original sound is a /fa/ from talker f103. When the entire fricative section is removed, it morphs into a robust /ba/.

3) Nasals: The two nasals /ma/ and /na/ share the common feature of a *nasal murmur* and differ from each other in the shape of F2 transition; specifically, /na/ has a prominent downward F2 transition while /ma/ does not. This is because the length of the vocal tract increases with /na/ as the tongue comes off the roof of the mouth, but stays the same length as the lips part; while for /ma/, the tongue remains on the floor of the mouth. Fig. 5 shows an example of /na/→/ma/ conversion. The original sound is a /na/ from talker f103; when the salient F2 transition is removed, it turns into a /ma/ for which some listeners can still prime /na/.

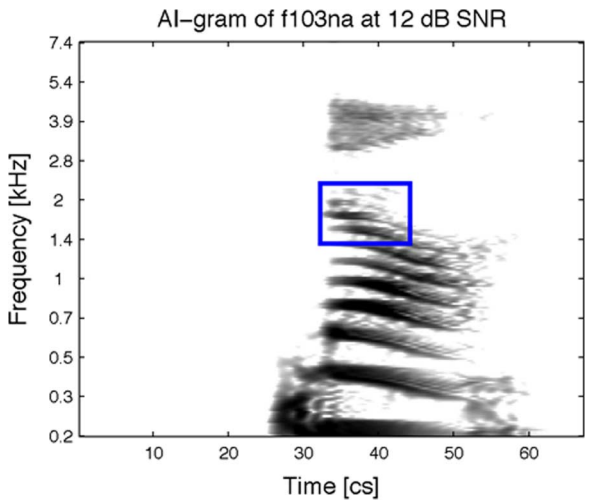

Fig. 5. AI-gram of /na/ from talker f103. Removing the downward F2 transition turns the /na/ into a /ma/. [Example: `na2ma`]. (1 cs = 0.01 s).

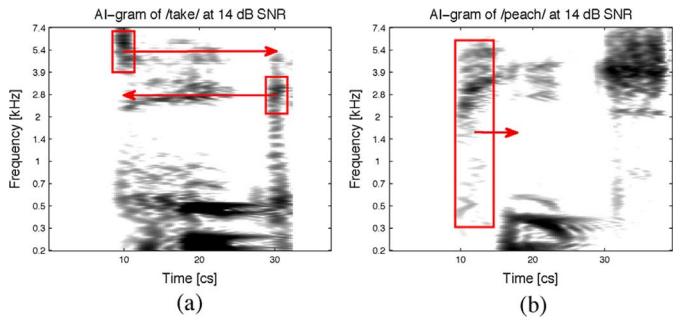

Fig. 6. Manipulation of words extracted from continuous speech. (a) a word /take/ morphs into /kate/ when the high-frequency /t/ cue is switched with the mid-frequency /k/ cue. [Example: `take2kate`] (b) a word /peach/ turns into /beach/ when the duration between the /p/ burst and the onset of sonorance is reduced from 60 to 0 ms. [Example: `peach2beach`]. (1 cs = 0.01 s). (a) /take/ → /kate/ (b) /peach/ → /beach/.

We have found that it is not always possible to manipulate the speech cue and turn a /ma/ into a convincing /na/, or vice versa, because the overall spectral patterns of the two sounds are quite different.

The very low-frequency "nasal murmur" though clearly audible does *not* seem to be a noise-robust cue used by listeners to label a sound as "nasal."

C. Words

A major difference between words and nonsense syllables is that words are meaningful. The semantic constraint can have a major impact on the perceptual integration of speech cues. Some researchers, especially those with linguistic background, do not believe that invariant cues exist for words and sentences. They seem to claim that speech perception is more about the interpretation of context information, rather than the detection and integration of perceptual cues.

In the previous section, we showed that the percept of nonsense CV syllables can be changed through the manipulation of speech cues. A key question is: *Does the same technique apply to words or sentences containing coarticulation and context?* To explore this question, we have chosen several words from our speech database and applied our speech-feature modification method. Fig. 6 shows two such examples, the words /take/ and /peach/, extracted from a sentence. As we see in Fig. 6(a), /t/ and /k/ are characterized respectively by a high-frequency burst at the beginning and a mid-frequency burst in the end. Switching the time location of the two cues turns the verb *take* into a perceived noun *Kate*. In Fig. 6(b), once the duration between the /p/ burst and the onset of sonorance is removed, /peach/ is reported as /beach/.

D. Sentences

The same technique of feature-based speech modification works for natural meaningful sentences, as shown in Fig. 7. Here we see the AI-gram of the sentence */she had your dark suit/* at 14-dB SNR (with phones labeled at the top). Removing the fricative cue around $t = 20$ cs (delete block 1 and 2) morphs the word /she/ into a /he/. Notice that the upper part of the /ʃa/ at 4–8 kHz (block 1) can then be used as the perceptual cue for an /s/; shifting it from $t = 20$ cs to $t = 55$ cs causes the word /had/ to morph to /has/. Next, we move the mid-frequency /k/ burst in the word /dark/ upward to 4 kHz, which converts the word /dark/ into /dart/. Finally, we change the /s/ cue in the word /suit/ to be a /ʃa/ cue by shifting it downward from 4–8 kHz to 2–4 kHz, morphs /suit/ into /shoot/. Thus, the modified sentence has become */he has your dart shoot/*. It is relatively easy to change the percept of most sounds once the consonant cues have been identified. Interestingly, meaningful sentences may easily be morphed into nonsense by modifying a single event. For example, we can turn the /d/ in /dark/ to a /b/ by zeroing out the frequency component above 1.4 kHz from 75 cs to 85 cs. The whole sentence then becomes */she has your bark suit/*.

The above examples of sentence modification clearly indicate that speech perception is critically dependent on specific speech cues. Context information becomes useful once the listener has decoded the speech cues. Specifically, while primes may be resolved by context, robust cues are not overpowered by such redundancy rendering context cues. A sentence may be described as having key words and accessory words. Similarly, the acoustic cues of continuous speech may be classified into two types: *critical* and *accessory* cues. The critical cues are defined as the irreplaceable units that are critical for perception of the sentence; the accessory cues refer to the redundant units recoverable from the critical cues and the associated context information.

Given *a priori* knowledge of perceptual cues, we have learned how to control the decoding of natural speech through the manipulation of speech cues in CV syllables, words, and sentences. This new understanding points to the feasibility of feature-based speech processing. In the next section, we will show that speech sounds can be made more robust to noise by manipulating the speech cues.

IV. Intelligibility of Consonants in Noise

We have demonstrated that speech perception is critically dependent on the detection of perceptual cues. When the dominant cue that defines a consonant is masked by noise under adversary environments, the conflicting cue may take effect and cause the listeners to report another consonant. The robustness of a consonant sound is determined by the strength of the dominant cue

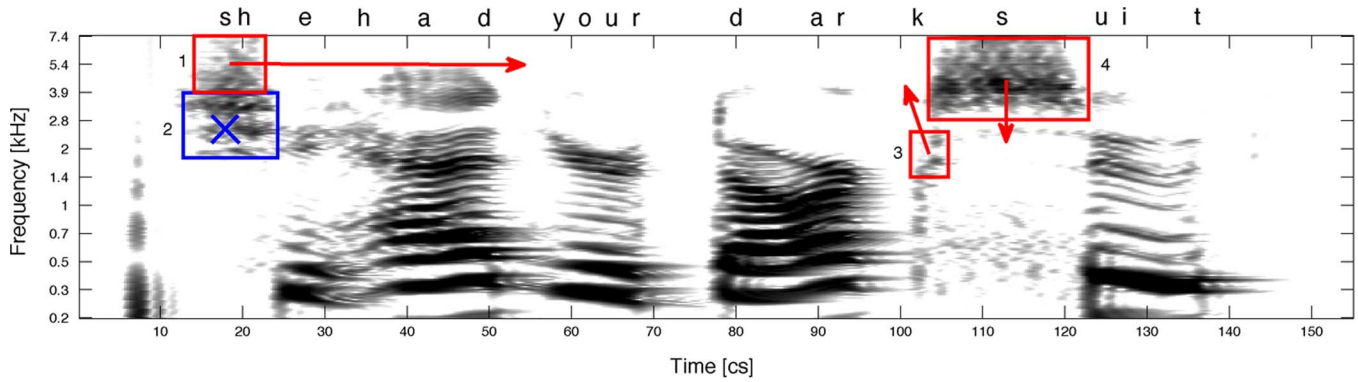

Fig. 7. Manipulation of speech cues converts a TIMIT sentence */she had your dark suit/* into a meaningful new sentence */he has your dart shoot/*. Step 1: convert /she/ into /he/ by removing the fricative part of /she/ (delete block 1 and 2); Step 2: to convert /had/ into /has/, a /s/ feature is created after /had/ by shifting the upper half of /ʃa/ feature (block 1) to $t = 55$ cs. Step 3: convert /dark/ into /dart/ by shifting the mid-frequency burst (block 3) upward. Step 4: convert /suit/ into a /shoot/ by shifting the /s/ cue (block 4) downward to 2–4 kHz. [Example: `she_had_your_dark_suit`]. (1 cs $= 0.01$ s).

TABLE I
CONFUSION MATRIX OF SPEECH PERCEPTION TEST ON STOP CONSONANTS

	-9 dB SNR (SWN)						-3 dB SNR (SWN)					
	pa	ta	ka	ba	da	ga	pa	ta	ka	ba	da	ga
pa	19	7	1	6	5	1	46	1	2	5		
ta	3	42	2		2	1		51	2			1
ka	12	**8**	**13**	3	5	3	6	**3**	**39**	4		1
ka $_{t\times 0}$	22	**5**	**4**	5		3	22	**4**	**16**	4	1	5
ka $_{t\times 0,\, k\times 2}$	7	**2**	**14**	2	1	6	6	**1**	**42**		1	4
ka $_{t\times 0,\, k\times 4}$	3	**1**	**27**	1	2	9	4	**2**	**42**		1	4
ba	4	1	3	8	7	5	8		1	31	6	1
da	5	11	2	3	25	1	1	1	1	1	44	3
ga	4	2	3	7	**16**	**12**	2	3	2	1	**16**	**26**
ga $_{d\times 0}$	4	3	2	8	**4**	**16**	1			8	**8**	**33**
ga $_{d\times 0,\, g\times 2}$	1	1	11	3	**10**	**20**				1	**5**	**42**
ga $_{d\times 0,\, g\times 4}$	1		9	4	**3**	**26**			1		**5**	**48**

$*$ $t, d \times 0$ means removing the interfering /ta/ or /da/ cue; $k, g \times N$ means amplifying /ka/ or /ga/ cue by a gain factor of N. The number of correct responses for /ka, ga/ and the number of confused responses, including both /ka/→/ta/ and /ga/→/da/, are in bold font.

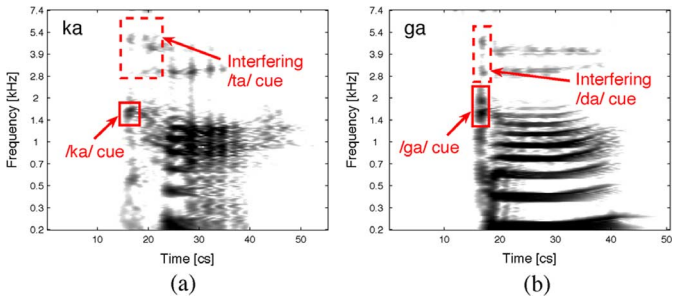

Fig. 8. Enhanced /ka/'s and /ga/'s were created by removing the high-frequency conflicting cues (dashed boxes) to promote /ta/→/ka/ responses and /ga/→/da/ confusions, and then boosting the mid-frequency bursts, critical for /ka/ and /ga/ identification. (1 cs $= 0.01$ s). (a) super /ka/ (b) super /ga/.

[55], [41]. To test the idea of improving speech intelligibility in noise by manipulating the speech cues, we conducted a small speech perception experiment on stop consonants /ka/ and /ga/ containing high-frequency conflicting cues for /ta/ and /da/. In order to improve the noise-robustness, and reduce the "bias" toward /ta/ and /da/, the utterances were modified so that the high-frequency conflicting cue was removed and the mid-frequency perceptual cue was amplified, as depicted in Fig. 8.

A. Methods

The speech stimuli include /pa, ta, ka, ba, da, ga/ and several enhanced "super" /ka/'s and "super" /ga/'s having the mid-frequency /ka/ and /ga/ cue amplified by 1 (0-dB gain), 2 (6-dB gain) and 4 (12-dB gain), respectively. The speech stimuli were chosen from the University of Pennsylvania's Linguistic Data Consortium (LDC) LDC2005S22 "Articulation Index Corpus" such that each nonsense CV syllable has six talkers, half male and half female. The speech stimuli were presented to both ears simultaneously under two SNR conditions, -9 and -3 dB SNR, using speech-weighted noise (SWN). The speech tokens were fully randomized across talkers, conditions, and consonants. Three normal hearing college students (male, age $<$ 30) participated in the study. All subjects were born in the U.S. with English being their first language. Each token (utterance$\times$SNR) was presented to each subject 18 times. A Matlab program controlled the procedure. Speech stimuli were presented to the listeners through Sennheisser HD 280-pro headphones. Following each presentation, subjects responded to the stimulus by clicking on the button labeled with the CV among sixteen choices /pa, ta, ka, fa, θa, sa, ʃa, ba, da, ga, va, ða, za, ʒa, ma, na/. In the case that the speech was totally unintelligible due to the noise, the subject was instructed to click a "Noise Only" button.

The speech stimuli were played at the most comfortable level (MCL) of the listeners, which was around 70-dB SPL.

B. Results

Results of the speech perception experiment indicate that boosting the mid-frequency /ka/ and /ga/ cue significantly increases the recognition scores in noise. Table I shows the confusion matrix of the speech test. Each row of the table represents the number of responses reported by the listeners when the sound on the left-most column is presented. At −9-dB SNR, removing the interfering high frequency cue from /ka/ reduces the /ta/ confusion from 8 (row 3, col 2) to 5 (row 4, col 2). However, due to the existence of a low-frequency burst below 1 kHz (indicative of /pa/), most subjects report the sound as a /pa/; hence, it also reduces the number of correct responses from 13 (row 3, col 3) to 4 (row 4, col 3). Enhancing the mid-frequency cue for the target sound by 12 dB increases the number of correct responses from 13 (row 3, col 3) for the original sound /ka/ to 27 (row 6, col 3) for the modified sound $\mathrm{ka}_{t\times 0, k\times 4}$; Similar results are observed for /ga/, for which the number of correct responses is 12 (row 9, col 6) for the original sound versus 27 (row 12, col 6) for the enhanced sound $\mathrm{ga}_{d\times 0,\, g\times 4}$. When the SNR increases from −9 to −3 dB, the advantage of feature manipulation is still large for /ga/ with the number of correct responses being 26 (row 9, col 12) for the original sound versus 48 (row 12, col 12) for the enhanced sound $(\mathrm{ga}_{d\times 0,\, g\times 4})$; the benefit of speech enhancement becomes minimal for /ka/ as the performance saturates.

V. Summary and Discussion

In order to identify the delicate features that characterize human speech perception, it is necessary to have a direct way of determining the cues from natural speech. Using the combined approach of AI-gram to predict speech audibility and 3DDS to measure the contribution of sub-speech component to perception, we have identified the perceptual cues for many initial consonants [7]. Based on this prior knowledge of the perceptual cues for natural speech [7], [48], [42], we have proposed a method for manipulating consonant sounds in the time–frequency domain and demonstrated the feasibility of feature-based speech processing. The following summarizes our major findings.

- Speech perception critically depends on the reception of perceptual cues. Through the manipulation of the conflicting cues, most often a tiny spot on the spectrogram, the target sound can be convincingly converted into a competing sound, as demonstrated by the selected examples in this paper.
- A speech sound can be made more robust to noise by boosting the defining speech cue, or the perceptual confusions can be reduced by removing the conflicting cue, directly demonstrating the potential of feature-based speech processing.
- The success of feature-based speech processing is largely dependent on the accuracy of identified speech cues. A slight change in a speech feature can lead to a huge difference in perception.

In this paper, all the examples of speech modification are created manually. A key element of a feature-based speech processing system is the feature detector. As shown earlier, these features are time–frequency features, so time–frequency detection and estimation theory may provide a method for automating this task. An early study [23] identified the formal connections between detection theory and quadratic (magnitude-based) time–frequency representations. Sayeed and Jones [57] discovered how to design such optimal detectors directly from training data as well as how to implement optimal approximations very efficiently using spectrograms. A recent study by Kim *et al.* [38] derived a method of calculating the reliability of each time–frequency region from clean speech signal. Despite these progresses, automating the detection of features from noisy speech remains a challenge.

Acknowledgment

The authors would like to thank M. Hasegawa-Johnson, A. Trevino, L. Pan, R. Serwy, R. Singh, A. Menon, and other members of the HSR group for constructive discussion and insightful comments.

References

[1] J. B. Allen, "Short time spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-25, no. 3, pp. 235–238, Jun. 1977.

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

[3] J. B. Allen, "Harvey Fletcher's role in the creation of communication acoustics," *J. Acoust. Soc. Amer.*, vol. 99, no. 4, pp. 1825–1839, 1996.

[4] J. B. Allen, "Consonant recognition and the articulation index," *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2212–2223, 2005.

[5] J. B. Allen, *Articulation and Intelligibility*. LaPorte, CO: Morgan and Claypool, 2005.

[6] 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, Nov. 1977.

[7] J. B. Allen and F. Li, "Speech perception and cochlear signal processing," *IEEE Signal Process. Mag.*, vol. 29, no. 4, pp. 117–123, Jul. 2009.

[8] M. C. Anzalone, L. Calandruccio, K. A. Doherty, and L. H. Carney, "Determination of the potential benefit of time–frequency gain manipulation," *Ear Hear.*, vol. 27, no. 5, pp. 480–492, 2006.

[9] R. Bentler and L. K. Chiou, "Digital noise reduction: An overview," *Trends Amplificat.*, vol. 102, pp. 67–82, 2006.

[10] S. E. Blumstein, K. N. Stevens, and G. N. Nigro, "Property detectors for bursts and transitions in speech perceptions," *J. Acoust. Soc. Amer.*, vol. 61, pp. 1301–1313, 1977.

[11] 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, pp. 1001–1017, 1979.

[12] S. E. Blumstein and K. N. Stevens, "Perceptual invariance and onset spectra for stop consonants in different vowel environments," *J. Acoust. Soc. Amer.*, vol. 67, pp. 648–266, 1980.

[13] R. Jakobson, C. Gunnar, M. Fant, and M. Halle, *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. Cambridge, MA: MIT Press, 1961, 39.

[14] 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.

[15] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *J. Acoust. Soc. Amer.*, vol. 102, no. 5, pp. 2892–2905, 1997.

[16] 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.

[17] H. W. Dudley, "The vocoder," *Bell Labs Rec.*, vol. 18, pp. 122–126, 1939.

[18] S. Dusan and L. R. Rabiner, "Can automatic speech recognition learn more from human speech perception?," in *Trends in Speech Technology*, C. Burileanu, Ed. Cluj Napoca, Romania: Romanian Academic Publisher, 2005, pp. 21–36.
[19] B. Delgutte, "Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers," *J. Acoust. Soc. Amer.*, vol. 63, no. 3, pp. 843–857, 1980.
[20] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1053–1064, 1994.
[21] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 5, pp. 2670–2680, 1994.
[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, 1984.
[23] P. Flandrin, "A time–frequency formulation of optimum detection," *IEEE Trans. Acoust. Speech and Signal Process.*, vol. 36, no. 9, pp. 1377–1384, Dec. 1988.
[24] G. Fant, *Speech Sounds and Features*. Cambridge, MA: MIT Press, 1973.
[25] J. L. Flanagan, *Speech Analysis Synthesis and Perception*. New York: Academic, 1965.
[26] H. Fletcher, "Speech and Hearing in Communication," in *The ASA Edition of Speech and Hearing in Communication*, J. B. Allen, Ed. New York: Acoust. Soc. of Amer., 1995, pp. A1–A34, 1–487.
[27] H. Fletcher and R. Galt, "Perception of speech and its relation to telephony," *J. Acoust. Soc. Amer.*, vol. 22, pp. 89–151, 1950.
[28] C. A. Fowler, "Segmentation of coarticulated speech in perception," *Percept. Psychophys.*, vol. 36, pp. 359–368, 1984.
[29] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, pp. 90–119, 1947.
[30] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Amer.*, vol. 80, pp. 1016–1025, 1986.
[31] S. Greenberg and T. Arai, "What are the essential cues for understanding spoken language?," *IEICE Trans. Inf. Syst.*, vol. E87-D, no. 5, pp. 90–119, 2004.
[32] D. D. Greenwood, "Critical bandwidth and the frequency coordinates of the basilar membrane," *J. Acoust. Soc. Amer.*, vol. 33, pp. 1344–1356, 1961.
[33] J. Heinz and K. Stevens, "On the perception of voiceless fricative consonants," *J. Acoust. Soc. Amer.*, vol. 33, pp. 589–596, 1961.
[34] H. Hermansky, "Should recongizers have ears?," *Speech Commun.*, vol. 25, pp. 3–27, 1998.
[35] J. T. Huang and M. Hasegawa-Johnson, "Maximum mutual information estimation with unlabeled data for phonetic classification," in *Proc. Interspeech*, 2008.
[36] G. W. Hughes and M. Halle, "Spectral properties of fricative consonants," *J. Acoust. Soc. Amer.*, vol. 28, no. 2, pp. 303–310, 1956.
[37] D. Kewley-Port, "Time-varying features as correlates of place of articulation in stop consonants," *J. Acoust. Soc. Amer.*, vol. 73, no. 1, pp. 322–335, 1983.
[38] G. Kim, Y. Lu, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1486–1494, 2009.
[39] K. D. Kryter, "Methods for the calculation and use of the articulation index," *J. Acoust. Soc. Amer.*, vol. 34, no. 11, pp. 1689–1697, 1962.
[40] H. Levitt, "Noise reduction in hearing aids: A review," *J. Rehab. Res. Develop.*, vol. 38, no. 1, pp. 111–121, 2001.
[41] F. Li, A. Menon, and J. B. Allen, "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," *J. Acoust. Soc. Amer.*, vol. 127, no. 4, pp. 2599–2610, 2010.
[42] F. Li, "Perceptual cues of consonant sounds and impact of sensorineural hearing loss on speech perception," Ph.D. dissertation, Univ. of Illinois at Urbana-Champaign, Urbana, 2009.
[43] A. Liberman, "Some results of research on speech perception," *J. Acoust. Soc. Amer.*, vol. 29, pp. 117–123, 1957.
[44] A. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, pp. 1–36, 1985.
[45] B. E. Lobdell, "Models of human phone transcription in noise based on intelligibility predictors," Ph.D. dissertation, Univ. of Illinois at Urbana-Champaign, Urbana, 2009.
[46] A. Male'cot, "Acoustic cues for nasal consonants: An experimental study involving a tape-splicing technique," *J. Acoust. Soc. Amer.*, vol. 32, no. 2, pp. 274–284, 1956.
[47] J. L. McClelland and J. L. Elman, "The trace model of speech perception," *Cognitive Psychol.*, vol. 18, pp. 1–86, 1986.
[48] A. Menon, F. Li, and J. B. Allen, "A psychoacoustic methodology to study perceptual cues of fricative consonants in natural speech," *J. Acoust. Soc. Amer.*, 2010, submitted for publication.
[49] 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.
[50] S. Phatak and J. B. Allen, "Consonant and vowel confusions in speech-weighted noise," *J. Acoust. Soc. Amer.*, vol. 121, no. 4, pp. 2312–2326, 2007.
[51] S. Phatak, A. Lovitt, and J. B. Allen, "Consonant confusions in white noise," *J. Acoust. Soc. Amer.*, vol. 124, no. 2, pp. 1220–1233, 2008.
[52] R. K. Potter, G. A. Kopp, and H. G. Kopp, *Visible Speech*. New York: Dover, 1966.
[53] L. R. Rabiner, "The power of speech," *Science*, vol. 301, pp. 1494–1495, 2003.
[54] A. Male'cot, "Place cues for nasal consonants with special reference to Catalan," *J. Acoust. Soc. Amer.*, vol. 73, no. 4, pp. 1346–1353, 1956.
[55] M. S. Regnier and J. B. Allen, "A method to identify noise-robust perceptual features: Application for consonant /t/," *J. Acoust. Soc. Amer.*, vol. 123, no. 5, pp. 2801–2814, 2008.
[56] R. Remez, P. Rubin, D. Pisoni, and T. Carrell, "Speech perception without traditional speech cues," *Science*, vol. 212, pp. 947–949, 1981.
[57] A. M. Sayeed and D. L. Jones, "Optimal detection using bilinear time–frequency and time-scale representations," *IEEE Trans. Signal Process.*, vol. 43, no. 12, pp. 2872–2883, Dec. 1995.
[58] 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.
[59] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, pp. 303–304, 1995.
[60] F.-G. Zeng, G. Stickney, Y. Kong, M. Vongphe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," *Proc. Natl. Acad. Sci.*, vol. 102, pp. 2293–2298, 2005.
[61] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," *J. Acoust. Soc. Amer.*, vol. 64, pp. 1358–1369, 1978.
[62] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1872–1891, 2002.
[63] R. Warren, "Perceptual restoration of missing speech sounds," *Science*, vol. 167, pp. 392–393, 1970.

Feipeng Li received the B.S. and M.S. degrees in electrical engineering from Wuhan University, Wuhan, China, in 1996 and 1999, respectively, and the Ph.D. degree from the Electrical and Computer Engineering Department, University of Illinois at Urbana-Champaign, Urbana, in 2009.

After graduation, he joined the National Remote Sensing Lab, Wuhan University, where he was a Research Scientist. Currently, he is a Postdoc Research Fellow at the Center for Hearing and Balance, Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD. His research interest is in human speech perception and speech processing.

Jont B. Allen (M'76–SM'79-F'85–LF'10) received the 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, Philadelphia, in 1968 and 1970, respectively.

After graduation in 1970, he joined Bell Laboratories, Murray Hill NJ, where he was in the Acoustics Research Department (from 1974 to 1997), as a Distinguished Member of Technical Staff. From 1997 to 2002, he was with the research division of the newly created AT&T Labs. Then in 2003, he joined the Department of Electrical Engineering, University of Illinois, as an Associate Professor, with tenure (2007).