A psychoacoustic method to find the perceptual cues of stop consonants in natural speech

Feipeng Li,[a] Anjali Menon, and Jont B. Allen
Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801

(Received 11 February 2009; revised 2 October 2009; accepted 31 December 2009)

Synthetic speech has been widely used in the study of speech cues. A serious disadvantage of this method is that it requires prior knowledge about the cues to be identified in order to synthesize the speech. Incomplete or inaccurate hypotheses about the cues often lead to speech sounds of low quality. In this research a psychoacoustic method, named three-dimensional deep search (3DDS), is developed to explore the perceptual cues of stop consonants from naturally produced speech. For a given sound, it measures the contribution of each subcomponent to perception by time truncating, highpass/lowpass filtering, or masking the speech with white 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 generally in agreement with the classical studies that stops are characterized by a short duration burst followed by a F2 transition, suggesting the effectiveness of the 3DDS method. However, it is also shown that /ba/ and /pa/ may have a wide band click as the dominant cue. F2 transition is not necessary for the perception of /ta/ and /ka/. Moreover, many stop consonants contain conflicting cues that are characteristic of competing sounds. The robustness of a consonant sound to noise is determined by the intensity of the dominant cue.
 [DOI: 10.1121/1.3295689]

PACS number(s): 43.71.Es [ADP] Pages: 2599–2610

I. INTRODUCTION

Speech sounds are characterized by time-varying spectral patterns called acoustic cues. When a speech wave propagates on the basilar membrane (BM), unique perceptual cues (named *events*), which define the basic units for speech perception, become resolved. The relationship between the acoustic cues and perceptual units has been a key research problem for speech perception (Fletcher and Galt, 1950; Allen, 1996, 2005a).

Bell Labs (1940): The first search for acoustic cues dates back to 1940s at Bell Laboratories, when Potter *et al.* (1966) began their *visible speech* project, with the goal of training the hearing-impaired to read spectrograms. Five normal hearing (NH) and one hearing-impaired (HI) listeners participated in the study. Following a series of lectures on the spectrograph and its use on isolated syllables and continuous speech, the subjects were successfully trained to "read" speech spectrographs. Even though the acoustic cues identified by visual inspection were not very accurate, this pioneering work laid a solid foundation for subsequent quantitative analysis.

Haskins Laboratories (1950): Cooper *et al.* (1952), along with other researchers at the Haskins Laboratories over the following decade, conducted a series of landmark studies on the acoustic cues of consonant sounds. A speech synthesis system, called the *Pattern Playback*, was created to convert a spectrograph into (low quality) speech. Based on the spectrographs of real speech, it was postulated that stop consonants are characterized by an initial burst and the following consonant-vowel transition. In this 1952 study (Cooper *et al.*, 1952), the authors investigated the effect of center frequencies of the burst and the second formant (F_2) transition, on the percept of unvoiced stop consonants, by using a set of "nonsense" consonant-vowel (CV) speech sounds synthesized from 12 bursts followed by seven F_2 formant frequencies. The subjects were instructed to identify the stimulus as /p/, /t/, or /k/ (a closed-set task). Results show that most people hear /t/ when the burst-frequency is higher than the F_2 frequency; when the two frequencies are close, most listeners report /k/; otherwise they hear /p/. In a following study (Delattre *et al.*, 1955), the authors dropped the burst and examined the effect of F_2 transition only on the percept of stop consonants. It was found that stimuli with rising F_2 transition were identified as /b/, those with F_2 emanating from 1.8 kHz were associated with /d/ and those with a falling transition were reported as /g/.

Follow-up studies (1960–1990): These early Haskins studies have had a major impact on the research of speech perception. Since then speech synthesis has become a standard method for feature analysis. It was used in the search of acoustic correlate for stops (Blumstein *et al.*, 1977), fricatives (Hughes and Halle, 1956; Heinz and Stevens, 1961), nasals (Malécot, 1973; Liberman, 1957; Recasens, 1983), as well as distinctive and articulatory features (Blumstein and Stevens, 1979, 1980; Stevens and Blumstein, 1978). Similar approach was taken by Remez *et al.* (1981) to generate highly unintelligible "sine-wave" speech, and then concluded that the traditional cues, such as bursts and transitions, are not required for speech perception.

[a]Author to whom correspondence should be addressed. Electronic mail: fli2@illinois.edu

The *status quo* is extremely confusing in that most people strongly believe that the stop consonants are defined by the bursts and transitions (Cooper *et al.*, 1952; Delattre *et al.*, 1955), yet still argue that modulation is the key to understand speech perception (Drullman *et al.*, 1994a, 1994b; Shannon *et al.*, 1995; Elliott and Theunissen, 2009). They failed to point out that the two views are actually in conflict.

The argument in favor of the speech synthesis method is that the features can be carefully controlled. However, the major disadvantage of synthetic speech is that it requires prior knowledge of the cues being sought. This incomplete and inaccurate knowledge about the acoustic cues has often led to synthetic speech of low quality, and it is common that such speech sounds are unnatural and barely intelligible, which by itself is a strong evidence that the critical cues for the perception of target speech sound are poorly represented. For those cases, an important question is "How close are the synthetic speech cues to those of natural speech?" Another key issue is the *variability* of natural speech, due to the talker (Hazan and Rosen, 1991), accent, masking noise, etc., most of which are well beyond the reach of the state-of-the-art speech synthesis technology. To answer questions such as "Why /ba/s from some of the talkers are confused with /va/, while others are confused with /ga/?" or "What makes one speech sound more robust to noise than another?," it is necessary to study the acoustic cues of naturally produced speech, not artificially synthesized speech.

This study explores a psychoacoustic method for isolating speech cues from natural CV speech. Rather than making assumptions about the cues to be identified, each natural speech utterance is modified by (1) adding noise of variable type and degree, (2) truncation of the speech from the onset, and (3) highpass and lowpass filtering the speech with variable cutoff frequencies. For each of these modifications, the identification of the sound is judged by a large panel of listeners. We then analyze the results to determine where in time, frequency, and at what signal to noise ratio (SNR) the speech identity has been masked and we characterize the confusion. In this way we triangulate on the location of the speech cues and the events, along the three independent dimensions. This procedure is thus called the three-dimensional deep search (3DDS) method.

A. Principle of the 3DDS

Speech sounds are characterized in three dimensions: time, frequency, and intensity. Event identification involves isolating the speech cues along these three dimensions. In the past studies, confusion test on nonsense syllables has long been used for the exploration of speech features. For example, Fletcher and colleagues investigated the contribution of different frequency bands to speech intelligibility using highpass and lowpass filtered CV and CVC syllables (Fletcher and Galt, 1950; French and Steinberg, 1947), resulting in the articulation index (AI) model. Furui (1986) examined the relationship between dynamic features and the identification of Japanese syllables modified by initial and final truncations. More often masking noise was used to study consonant (Miller and Nicely, 1955; Wang and Bilger, 1973) and vowel (Phatak and Allen, 2007) recognition. Régnier and Allen (2008) successfully combined the results of time truncation and noise-masking experiments for the identification of /ta/ events. However, it has remained unclear how many speech cues could be extracted from real speech by these methods. In fact, there seems to be high skepticism within the speech research community as the general utility of any proposed methods.

In the present investigation, we have integrated the three types of tests, thus proposing the "3DDS" method for exploring the events of consonants from natural speech. To evaluate the acoustic cues along the three dimensions, speech sounds are truncated in time, highpass/lowpass filtered, and masked with white noise, as illustrated in Fig. 1, and then presented to NH listeners for identification. Small close set tasks are avoided because of their inherent bias.

Imagine that an acoustic cue, critical for speech perception, has been removed or masked. Would this degrade the speech sound and reduce the recognition score significantly? For the sound /t/, Régnier and Allen (2008) answered this question: The /t/ event is entirely due to a single short ≈20 ms burst of energy, between 4 and 8 kHz. To estimate the importance of individual speech perception events for sounds other than /t/, the 3D approach requires three independent experiments for each CV utterance. The *first* experiment 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 sound and its duration. The *second* experiment divides the fullband into multiple bands of equal length along the BM and measures the score in these different frequency bands. Once the time-frequency coordinates of the event have been identified, a *third* experiment assesses the strength of the speech event by masking the speech at various signal-to-noise ratios. This experiment determines the score as a function of the SNR, as required for an AI calculation. To reduce the length of the experiments, the three dimensions, i.e., time, frequency, and intensity, are independently measured. The identified events have been further verified by a special software package designed for the manipulation of acoustic cues (Allen and Li, 2009) based on the short-time Fourier transform (Allen, 1977; Allen and Rabiner, 1977).

In order to understand continuous speech, it is necessary to first identify the acoustic correlates of the individual phonemes, for which the movement of the articulators are more easily interpretable (Fant, 1973). For this reason, as in the 1950 Haskins studies, we first look at the normal events of individual consonants in isolated CV syllables. The interaction between the events in continuous speech must be addressed in future studies. Finally, the 3DDS method has been successfully applied to all of the 16 Miller–Nicely consonants followed by three vowels /i, ɑ, u/, but for both space and pedagogical reasons, the discussion here has been limited to the six stop consonants /p, t, k, b, d, g/ preceding vowel /a/. Brief summaries of portion of this study have been presented in Allen *et al.*, 2009 and Allen and Li, 2009. The work was done as part of the Ph.D. thesis of Li.

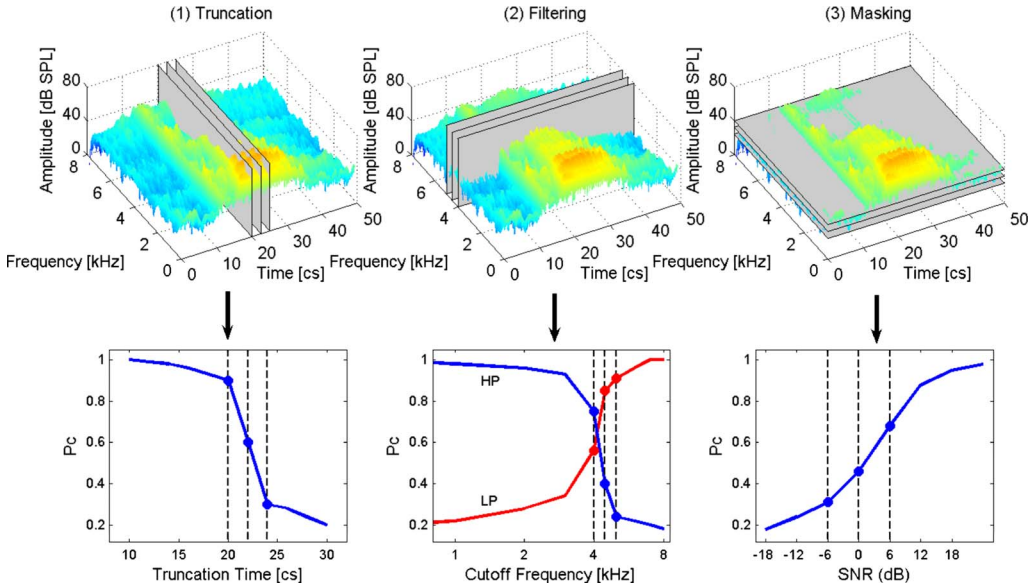


FIG. 1. (Color online) The 3D approach for the identification of acoustic cues: (1) to isolate the cue along the time, speech sounds are truncated in time from the onset with a step size of 5, 10, or 20 ms, depending on the duration and type of consonant; (2) to locate the cue along the frequency axis, speech sounds are highpass and lowpass filtered before being presented to normal hearing listeners; and (3) to measure the strength of the cue, speech sounds are masked by white noise of various signal-to-noise ratio. The three plots on the top row illustrate how the speech sound is processed. Typical correspondent recognition scores are depicted in the plots on the bottom row.

II. METHODS

The detail of the time-truncation (TR07), highpass/lowpass filtering (HL07), and noise-masking "Miller–Nicely (2005)" (MN05) experiments are described below. Each abbreviation gives the experiment type followed by the year the experiment was executed. An analysis of the MN05 experiment (also known as experiment: MN16R, publication: PLA08) has since been previously published (Phatak *et al.*, 2008), and results of HL07 can be found in Allen and Li (2009).

A. Subjects

In all, 61 listeners were enrolled in the study, of which 19 subjects participated in HL07, 19 in TR07. One subject participated in both of the experiments. The rest of the 24 subjects were assigned to experiment MN05 (Phatak *et al.*, 2008). The large majority of the listeners were undergraduate students, while the remaining were mothers of teenagers. No subject was older than 40 years, and all self-reported no history of speech or hearing disorder. All listeners spoke fluent English, with some having slight regional accents. Except for two listeners, all the subjects were born in the United States with their first language (L1) being English. The subjects were paid for their participation. University IRB approval was obtained.

B. Speech stimuli

A significant characteristic of natural speech is the variability of the acoustic cues. Thus we designed the experiment by manually selecting six different utterances per CV consonant based on the criterion that the samples be representative of the corpus. This decision was based on the scores from MN05. In retrospect this was a minor tactical mistake, as the most is learned from the sounds having 100% scores in quiet.

Six talkers saying the 16 Miller and Nicely (1955) (MN55) CVs /pa, ta, ka, fa, θa, sa, ʃa, ba, da, ga, va, ða, za, ʒa, ma, na/ were chosen from the University of Pennsylvania's Linguistic Data Consortium (LDC) LDC2005S22 "Articulation Index Corpus," which is used as the common speech source for the three experiments. This corpus is described in some detail by Fousek *et al.* (2004). Briefly, the speech sounds were sampled at 16 kHz using a 16 bit analog to digital converter. Each CV was spoken by 20 talkers of both genders. Experiment MN05 uses 18 talkers $\times 16$ consonants. For the other two experiments (TR07 and HL07), six talkers, half male and half female, each saying each of the 16 MN55 consonants, were manually chosen for the test. These 96 (6 talkers $\times 16$ consonants) utterances were selected such that they were representative of the speech material in terms of confusion patterns and articulation scores based on the results of two earlier speech perception experiments (Phatak and Allen, 2007; Phatak *et al.*, 2008).

The speech sounds were presented diotically (same sounds to both ears) through a Sennheisser "HD 280 Pro" headphone, at each listener's "most comfortable level" (MCL) (i.e., between 75 and 80 dB sound pressure level, and calibrated using a continuous 1 kHz tone into a homemade 3 cm^3 flat-plate coupler, as measured with a Radio Shack sound level meter). All experiments were conducted in a single-walled IAC sound-proof booth. Typically the room holding the booth had the door shut and people in the room were instructed to speak softly, so that their speech would not distract with the subject in the booth.

C. Conditions

Three independent experiments were performed, denoted TR07, HL07, and MN05. All three experiments included a common condition of fullband speech at 12 dB SNR in white noise, as a control.

Experiment TR07 evaluates the temporal property of the events. Truncation starts from just before the beginning of the utterance and stops at the end of the consonant. The starting, stopping, and truncation times were manually chosen, such that the duration of the consonant was divided into nonoverlapping consecutive intervals of 5, 10, and 20 ms. An adaptive scheme was applied for the calculation of the sample points. The basic idea was to assign more points where the speech changed rapidly, and fewer points where the speech was in a steady condition, in a manner consistent with the findings of Furui (1986). Starting from the end of the consonant, near the consonant-vowel transition, 8 frames of 5 ms were allocated, followed by 12 frames of 10 ms, and as many 20 ms frames, as needed, until the entire interval of the consonant was covered. To make the truncated speech sounds more natural and to remove possible minor onset truncation artifacts, white noise was used to mask the speech stimuli, at a SNR of 12 dB.

Experiment HL07 investigates the frequency properties of the events (Li and Allen, 2009). Nineteen filtering conditions, including one full-band (250–8000 Hz), nine highpass, and nine lowpass conditions, were included. The cutoff frequencies were calculated using Greenwood function (Greenwood, 1990) so that the full-band frequency range was divided into 12 bands, each having equal length along the basilar membrane. The highpass cutoff frequencies were 6185, 4775, 3678, 2826, 2164, 1649, 1250, 939, and 697 Hz, with an upper limit of 8000 Hz. The lowpass cutoff frequencies were 3678, 2826, 2164, 1649, 1250, 939, 697, 509, and 363 Hz, with the lower limit being fixed at 250 Hz. Note that the highpass and lowpass filtering share the same cutoff frequencies over the middle range. The filters were implemented in MATLAB© (The Mathworks Inc.) via a sixth order elliptical filter, with a stop band of 60 dB. White noise having a 12 dB SNR was added, to assure that the modified speech has no audible out-of-band component.

Experiment MN05 assesses the strength of the event in terms of noise robust speech cues, under adverse conditions of high noise. Besides the quiet condition, speech sounds were masked at eight different SNRs: −21, −18, −15, −12, −6, 0, 6, and 12 dB, using white noise. The results reported here are a subset of the Phatak and Allen (2007) study, which provides the full details.

D. Procedures

The three experiments employed similar procedures. A mandatory practice session was given to each subject at the beginning of each experiment. The stimuli were fully randomized across all variables when presented to the subjects, with one important exception to this rule being MN05 where effort was taken to match the experimental conditions of Miller and Nicely (1955) as closely as possible (Phatak *et al.*, 2008). Following each presentation, subjects responded to the stimuli by clicking on a button labeled with the CV that they heard. In case the speech was completely masked by the noise, the subject was instructed to click a "noise only" button. If the presented token did not sound like any of the 16 consonants, the subject were told to either guess 1 of the 16 sounds, or click the noise only button. To prevent fatigue, listeners were told to take frequent breaks, or break whenever they feel tired. Subjects were allowed to play each token for up to three times before making their decision, after which the sample was placed at the end of the list. Three different MATLAB programs were used for the control of the three procedures. The audio was played using a SoundBlaster 24 bit sound card in a standard PC Intel computer, running Ubuntu Linux.

III. MODELING SPEECH RECEPTION

The cochlea decomposes each sound through an array of overlapping nonlinear (compressive), narrow-band filters, splayed out along the BM, with the base and apex of BM being tuned to 20 kHz and 20 Hz, respectively (Allen, 2008). Once a speech sound reaches the inner ear, it is represented by a time-varying response pattern along the BM, of which some of the subcomponents contribute to speech recognition, while others do not. Many components are masked by the highly nonlinear forward spread (Duifhuis, 1980; Harris and Dallos, 1979; Delgutte, 1980) and upward spread of masking (Allen, 2008). The purpose of event identification is to isolate the specific parts of the psychoacoustic representation that are required for each consonant's identification (Régnier and Allen, 2008).

To better understand how speech sounds are represented on the BM, the AI-gram (see Appendix A) is used. This construction is a signal processing auditory model tool to visualize audible speech components (Lobdell, 2006, 2008; Régnier and Allen, 2008). The AI-gram is thus called, due to its estimation of the speech audibility via Fletcher's AI model of speech perception (Allen, 1994, 1996), was first published by Allen (2008), and is a linear Fletcher-like critical band filter-bank cochlear simulation. Integration of the AI-gram over frequency and time results in the AI measure.

A. A preliminary analysis of the raw data

The experimental results of TR07, HL07, and MN05 are presented as *confusion patterns* (CPs), which display the probabilities of all possible responses (the target and competing sounds), as a function of the experimental conditions, i.e., truncation time, cutoff frequency, and signal-to-noise ratio.

Notation: Let $c_{x|y}$ denote the probability of hearing consonant /x/ given consonant /y/. When the speech is truncated to time t_n, the score is denoted $c^T_{x|y}(t_n)$. The scores of the lowpass and highpass experiments at cutoff frequency f_k are indicated as $c^L_{x|y}(f_k)$ and $c^H_{x|y}(f_k)$. Finally, the score of the masking experiment as a function of signal-to-noise ratio is denoted $c^M_{x|y}(\mathrm{SNR}_k)$.

The specific example of Fig. 2 is helpful to explain the 3DDS method and to show how speech perception is affected by the events. It depicts the CPs of a /ka/ produced by

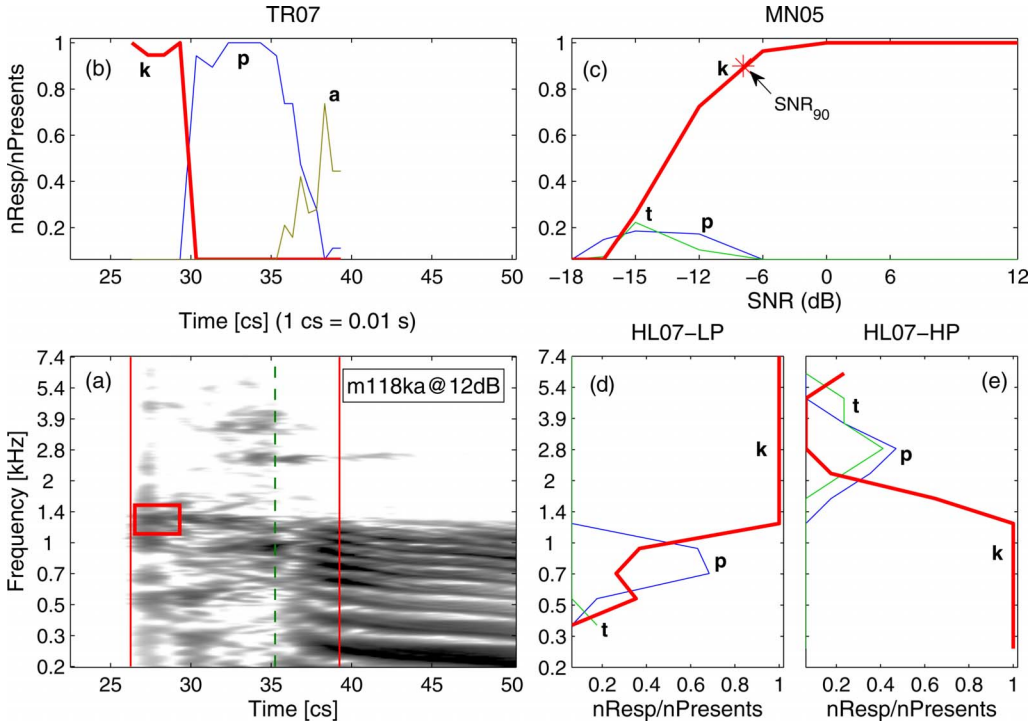


FIG. 2. (Color online) Various CPs of /ka/ spoken by talker m118 under various experimental conditions. (a) AI-gram at 12 dB SNR. The left and right vertical lines denote the start and end times for truncation. The middle line denotes the time of voice (sonorant) onset. (b) The temporal truncation CP as a function of truncation time from experiment TR07. (c) CP as a function of SNR for experiment MN05. The SNR_{90} point indicated by $*$ is at -8 dB SNR. Finally, the (d) low and (e) high CPs as a function of cutoff frequency for HL07. The text provides further details.

talker m118 (utterance m118_ka). The lower-left panel (a) shows the AI-gram. The results of the time-truncation experiment (TR07) is given in upper-left panel (b), lowpass/highpass of HL07 in lower-right panels [(d) and (e)], and MN05 in upper-right panel (c). To facilitate the integration of the three experiments, the AI-gram and the three scores are aligned in time t (in centiseconds) and frequency (along the cochlear place axis, in kilohertz) and thus depicted in a compact and uniform manner. Note that 1 cs=10 ms=0.01 s.

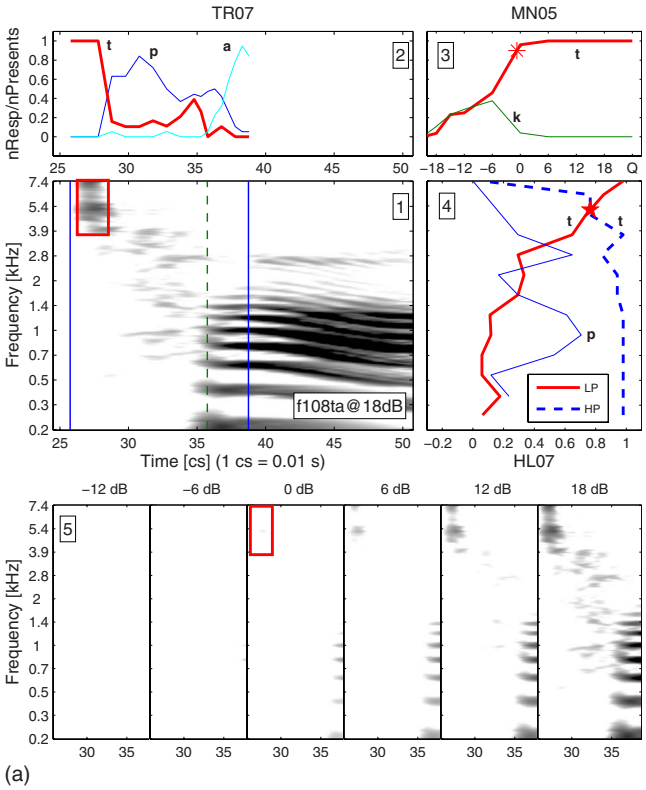
The CP of TR07 [Fig. 2(a)] shows that the probability of hearing /ka/ is 100% for $t \leq 26$ cs, where no speech component are removed. At 29 cs where the /ka/ burst has been completely truncated, the score for /ka/ drops sharply to 0%, within a span of 1 cs. For truncation times greater than 29 cs, only the transition region is heard, and 100% of the listeners report hearing a /pa/. Once the transition region is truncated ($t > 35$ cs), listeners report hearing only the vowel /a/.

A related conversion occurs in the lowpass and highpass experiment HL07 for /ka/ [Figs. 2(d) and 2(e)], in which both the lowpass score $c^L_{k|k}$ and highpass score $c^H_{k|k}$ abruptly plunge from 100% to less than 10% at a cutoff frequency of f_k=1.4 kHz, thereby precisely defining the frequency location of the /ka/ cue. For the lowpass case, listeners reported a morphing from /ka/ to /pa/ with score $c^L_{p|k}$ reaching 70% at 0.7 kHz, and for the highpass case, /ka/ morphed to /ta/, but only at the $c^H_{t|k}$=0.4 (40%) level. To reduce clutter, the remaining confusions are not shown.

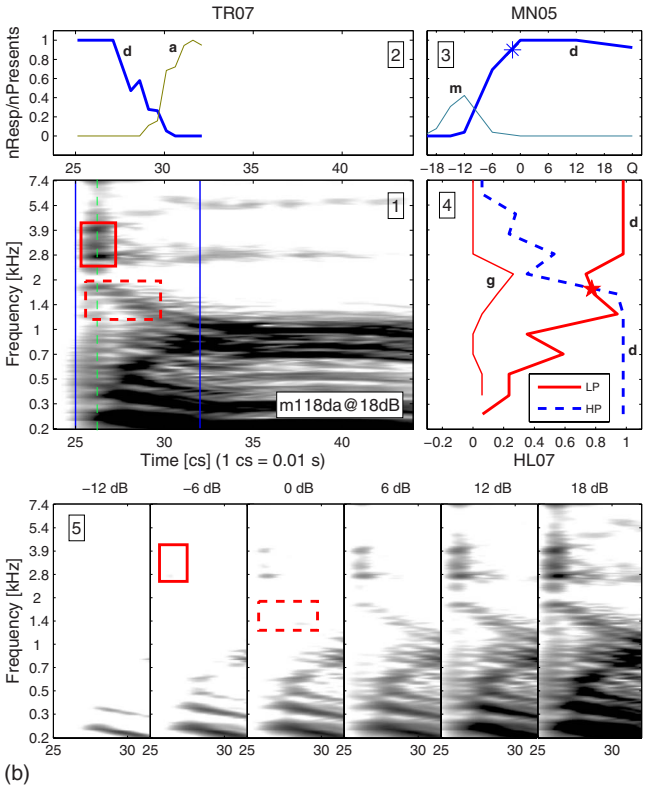
The MN05 masking data [Fig. 2(d)] show a fourth CP. When the masker level increases from quiet to 0 dB SNR, the recognition score of /ka/ is close to 1 (i.e., 100%), signifying the presence of a robust cue. At SNR_{90}=−8 dB, the score sharply drops to chance performance where it is confused with /t/ and /p/.

IV. RESULTS

In this section we demonstrate how the events of stop consonants are identified by applying the 3DDS method. Again the results from the three experiments are arranged in an abbreviated and even more compact form. In Fig. 3(a) panel 1 (middle left) shows the AI-gram of the speech sound at 18 dB SNR. Each event hypothesis is highlighted by a rectangular box. The middle vertical dashed line denotes the voice-onset time, while the two vertical solid lines on either side of the dashed line denote the starting and ending points for the time-truncation experiment (TR07). Directly above the AI-gram, panel 2 shows the scores from TR07, while to the right, panel 4 shows the scores from HL07. Panel 3 (upper right) depicts the scores from experiment MN05. The CP functions are plotted as solid (lowpass) or dashed (highpass) curves, with competing sound scores with a single letter identifier next to each curve. The $*$ in panel 3 indicates SNR_{90}=−2 dB where the listeners just begin to confuse the sound in MN05, while the $\star$ in panel 4 indicates the intersection point (1.3 kHz) of the highpass and lowpass scores. The six small figures along the bottom show partial AI-grams of the consonant region, delimited in panel 1 by the solid lines, at −12, −6, 0, 6, 12, and 18 dB SNRs. A box in any of the seven AI-grams of panel 1 or 5 indicates a hypothetical event region, and for panel 5 indicates its audible threshold predicted by the AI-gram model.



(a)



(b)

FIG. 3. (Color online) Hypothetical events for high-frequency stop consonants /ta/ and /da/. The multiple panels in each subfigure are [1] AI-gram. A dashed vertical line labels the onset of voicing (sonorance), indicating the start of the vowel. The solid and dashed boxes indicate the dominant and minor events, respectively. [2] CPs as a function of truncation time t_n. [3] CPs as a function of SNR_k. [4] CPs as a function of cutoff frequency f_k. [5] AI-grams of the consonant region (defined by the solid vertical lines on panel [1]) at −12, −6, 0, 6, 12, and 18 dB SNRs.

In Secs. IV A–IV C, we study the six stop consonants /t/, /d/, /k/, /g/, /p/, and /b/ followed by vowel /a/ as in "father." For each consonant, the six token utterances were analyzed by the members of our research group, and the most representative example was subjectively chosen to be presented, since it is impossible to publish all the data. An extensive attempt was made to automatically quantify the measures objectively; however, eventually this approach was abandoned, as it seriously obscured the raw data. Thus for this initial presentation of the 3DDS method, we decided to stick with a raw data presentation.

A. /ta/ and /da/

/ta/: Results of the three experiments (TR07, HL07, and MN05) clearly indicates that the /ta/ event [refer to Fig. 3(a) for a /ta/ from talker f105] is a high-frequency burst above 3 kHz, 1.5 cs in duration and 5–7 cs prior to the vowel. Panel [1] shows the AI-gram of the sound at 18 dB SNR in white noise with the hypothetical /ta/ event being highlighted by a rectangular frame. Above, panel [2] depicts the results of the time-truncation experiment. When the burst is completely removed at 28 cs, the score for the time-truncated /t/ drops dramatically from 1 to chance, and listeners start reporting /pa/, suggesting that the high-frequency burst is critical for /ta/ perception. This is in agreement with the highpass and lowpass data of panel [4]. Once the high-frequency burst has been removed by the lowpass filtering (solid curve), the /ta/ score $c^L_{t|t}$ drops dramatically and the confusion with /pa/ increases significantly. The intersection of the highpass and the lowpass perceptual scores (indicated by the ⋆) is ≈5 kHz, consistent with a high-frequency burst dominant cue. These results are then confirmed by the noise-masking experiment. From the AI-grams in panel [5], we see that the high-frequency burst becomes inaudible when the SNR is lower than 0 dB, as a consequence, the recognition score drops sharply at −1 dB SNR (labeled by a ∗ in panel [3]), proving that the perception of /ta/ is dominated by the high-frequency burst.

Of the six /ta/ sounds, five morph to /pa/ once the /ta/ burst was truncated, while one morphs to /ka/ (m112ta). For this particular sound, it is seen that the /ta/ burst precedes the vowel only by around 2 cs as opposed to 5–7 cs as is the case for a typical /ta/. This timing cue is especially important for the perception of /pa/, as we will discuss later in Sec. IV C.

/da/: Consonant /da/ [Fig. 3(b)] is the voiced counterpart of /ta/. It is characterized by a high-frequency burst above 4 kHz and a F_2 transition near 1.5 kHz, as shown in panel [1]. Truncation of the high-frequency burst (panel [2]) leads to an immediate drop in the score of $c^T_{d|d}$ from 100% at 27 cs to about 70% at 27.5 cs. The recognition score keeps decreasing until the F_2 transition is removed completely at 30 cs. From the highpass and lowpass data (panel [4]), it is seem that subjects need to hear both the F_2 transition and the high-frequency burst to get a full score of 100%, Lack of the burst usually leads to the /da/ → /ga/ confusion, as shown by the lowpass confusion of $c^L_{g|d}$=30% at f_c=2 kHz (solid curve labeled "g" in panel [4]), meaning that both the high-frequency burst and F_2 transition are important for the iden-

tification of a high quality /da/. This is confirmed by the results of the noise-masking experiment. From the AI-grams (panel 5), the F_2 transition becomes masked by noise at 0 dB SNR; accordingly the /da/ score $c^M_{d|d}$ in panel 3 drops quickly at the same SNR. When the remnant of the high-frequency burst is finally gone at −6 dB SNR, the /da/ score $c^M_{d|d}$ decreases even faster, until $c^M_{d|d}=c^M_{m|d}$ at −10 dB SNR; namely, the /d/ and /m/ scores are equal.

Some of the /da/'s are much more robust to noise than others. For example, the SNR_{90}, defined as the SNR where the listeners begin to lose the sound ($P_c=0.90$), is −6 dB for /da/-m104, and +12 dB for /da/-m111. The variability over the six utterances is impressive, yet the story seems totally consistent with the requirement that both the burst and the F_2 onset need to be heard.

B. /ka/ and /ga/

/ka/: Analysis of Fig. 4(a) reveals that the event of /ka/ is a mid-frequency burst around 1.6 kHz, articulated 5–7 cs before the vowel, as highlighted by the rectangular boxes in panels 1 and 5. The truncation data in panel 2 show that once the mid-frequency burst is truncated at 16.5 cs, the recognition score $c^T_{k|k}$ jumps from 100% to chance level within 1–2 cs. At the same time, most listeners report /pa/. The highpass score $c^H_{k|k}$ and the lowpass score $c^L_{k|k}$ (panel 4) both label 1.6 kHz with a sharp intersection, suggesting that the perception of /ka/ is dominated by the mid-frequency burst. Based on the AI-grams (panel 5), the 1.6 kHz burst is just above its detection threshold at 0 dB SNR, accordingly the recognition score of /ka/ $c^M_{k|k}$ (panel 3) drops dramatically below 0 dB SNR. Thus the results of the three experiments seem in perfect agreement in identifying a $\frac{1}{2}$ octave wide burst around 1.6 kHz as the single dominant cue of /ka/.

The identified /ka/ burst is consistent across all talkers. Four of the six /ka/ sounds morph to /pa/ once the /ka/ burst was truncated. Two have no morphs, remaining a very weak /ka/ (m114ka, f119ka).

/ga/: Consonant /ga/ [Fig. 4(b)] is the voiced counterpart of /ka/. Like /ka/, it is represented by a $\frac{1}{2}$ octave burst from 1.4 to 2 kHz, immediately followed by a F_2 transition between 1 and 2 kHz, all highlighted with boxes in panel 1. According to the truncation data (panel 2), the recognition score of /ga/ $c^T_{g|g}$ starts to drop once the mid-frequency burst is truncated beyond 21 cs. By 23 cs, a /ga/ → /da/ confusion appears with $c^T_{d|g}=40\%$. The highpass and lowpass scores (panel 4) fully overlap at 1.6 kHz, where both show a sharp decrease of more than 60%, consistent with the time-truncation data for /ga/. Based on the AI-grams in panel 5, the F_2 transition is masked by 0 dB SNR, while $SNR_{90}\approx-2$ dB, as labeled by a * in panel 3. When the mid-frequency burst is fully masked at −6 dB SNR, /ga/ becomes confused with /da/, suggesting that the perception of /ga/ is dominated by the mid-frequency burst.

All six /ga/ sounds have well defined bursts between 1.4 and 2 kHz. Most of the /ga/s (m111, f119, m104, and m112) have a perfect score of $c^M_{g|g}=100\%$ at 0 dB SNR. The other two /ga/s (f109 and f108) are a few dB weaker.

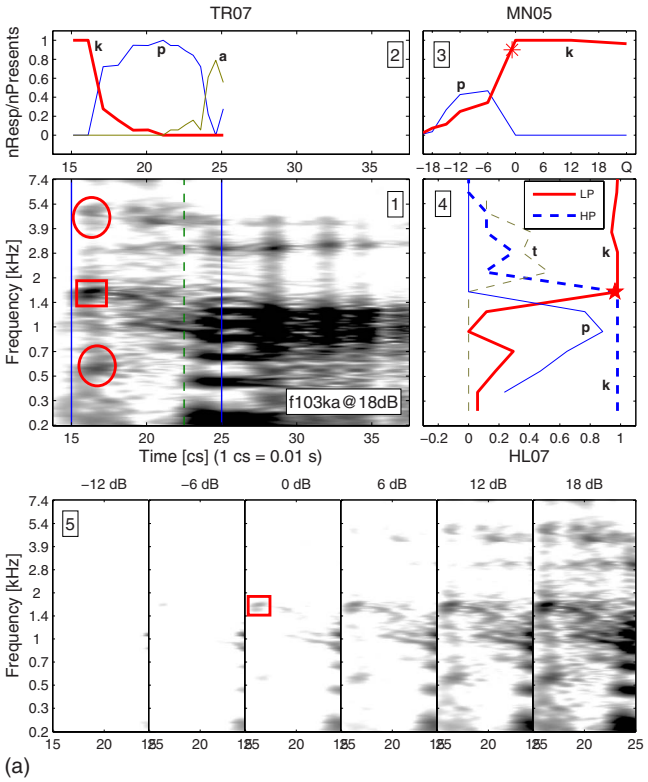
It is interesting to note that these two mid-frequency sounds all have conflicting cues that are characteristic of competing sounds. For example, the /ka/ sound [Fig. 4(a)] also contains a high-frequency burst around 5 kHz and a low-frequency burst around 0.5 kHz that that could be used as a perception cue of /ta/ and /pa/, respectively. As a consequence, listeners hear /ta/ when the highpass cutoff frequency is higher than the upper limit of /ka/ burst (2 kHz). In the lowpass experiment, people hear /pa/ when the lowpass cutoff frequency is smaller than 1.2 kHz, the lower limit of /ka/ cue. Similarly the /ga/ also contains a high-frequency burst above 4 kHz that promotes the confusion of /da/.

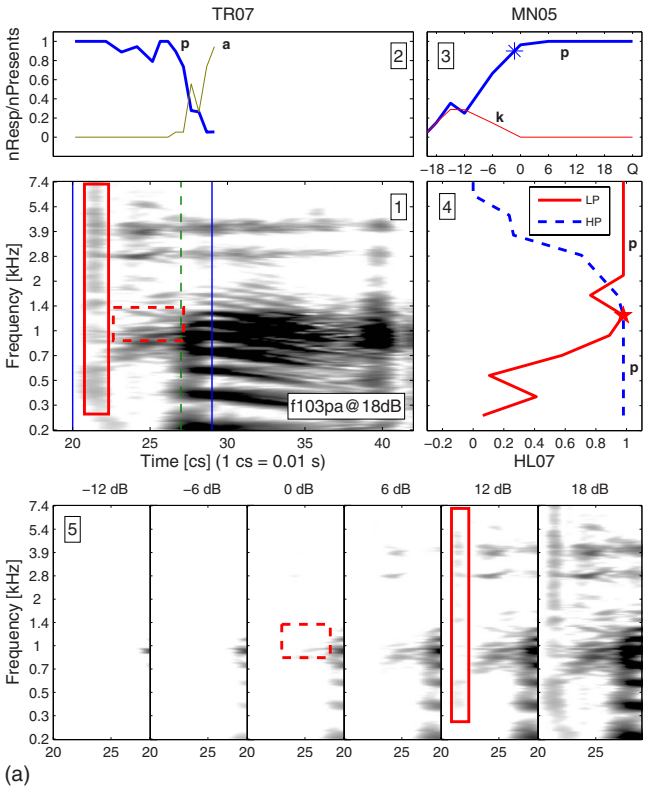
C. /pa/ and /ba/

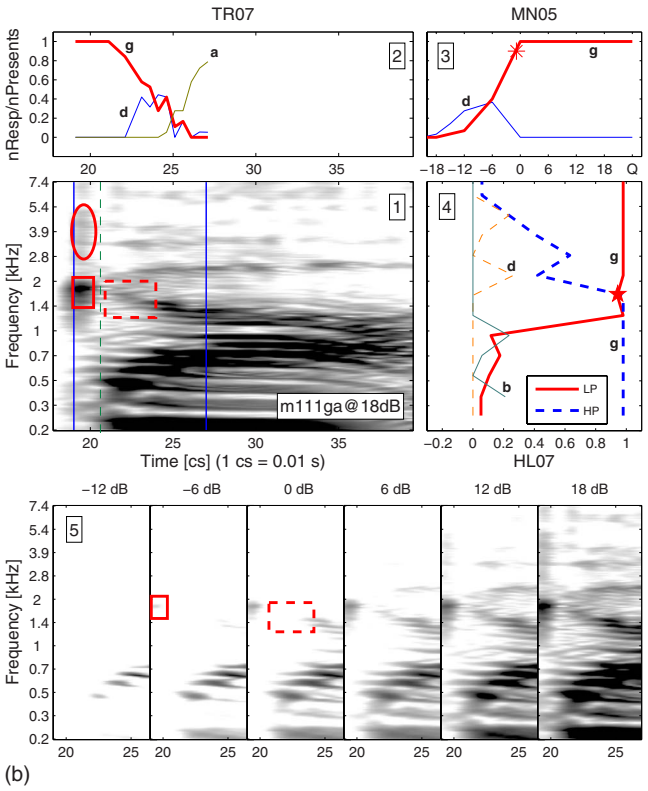
/pa/: The AI-gram in Fig. 5(a) 1 for /pa/ spoken by female talker f103 reveals that there could be two different potential events: (1) a wide band click running from 0.3 to 7.4 kHz, maskable by white noise at 6 dB SNR and (2) a formant resonance at 1–1.4 kHz, maskable by white noise at 0 dB SNR. Panel 2 shows the truncated /p/ score $c^T_{p|p}(t_n)$. It starts at 100%. Once the wide band click spanning 0.3–7 kHz is truncated at ≈22–23 cs, the score drops out of saturation. Once the transition is removed at 27 cs it further drops to chance (1/16). The lowpass and highpass scores (panel 4) start at 100% at each end of the spectrum, and drop around the intersection point between 1.4 and 2 kHz. This broad intersection (indicated by a ⋆) appears to be a clear indicator of the center frequency of the dominant perceptual cue, which is at $F_2\approx0.7-1.0$ kHz and before 22–26 cs. The recognition score of noise-masking experiment (panel 3) drops dramatically at $SNR_{90}\approx=-1$ dB SNR (denoted by a *). From the six AI-grams (panel 5), we can see that the predicted audible threshold for the F_2 transition is at 0 dB SNR, the same as SNR_{90} (*) in panel 3 where the listeners just begin to lose the sound. Thus both the wide band click and the F_2 onset contribute to the perception of /pa/.

Stop consonant /pa/ is characterized as having a wide band click, as seen in this /pa/ example, but not in the five others we have studied. For most /pa/'s, the wide band click diminishes into a low-frequency burst. When the click is partially removed by filtering, the score remains at 100% as long as the F_2 region is audible. The click appears to contribute to the overall quality of /pa/. The 3D displays of other five /pa/s are in basic agreement with that of Fig. 5(a), with the main difference being the existence of the wideband burst at 22 cs for f103, and slightly different highpass and lowpass intersection frequencies, ranging from 0.7 to 1.4 kHz. The required duration of the F_2 energy before the onset of voicing (around 3–5 cs) is consistently critical for all the /pa/ utterances.

/ba/: Identifying the perceptual events for /ba/ are perhaps the most difficult of the six stops. For the 3DDS method to work well, 100% scores in quiet are required. Among the six /ba/ sounds, only the one in Fig. 5(b) has 100% scores at 12 dB SNRs and above. Based on the analysis of the AI-gram of Fig. 5(b), the potential features for /ba/ is a wide band click in the range of 0.3–4.5 kHz. Once the wide band click is completely truncated by $t_n=27$ cs, the /ba/ score $c^T_{b|b}$ [Fig. 5(b) 2] drops dramatically from 80% to chance, at the







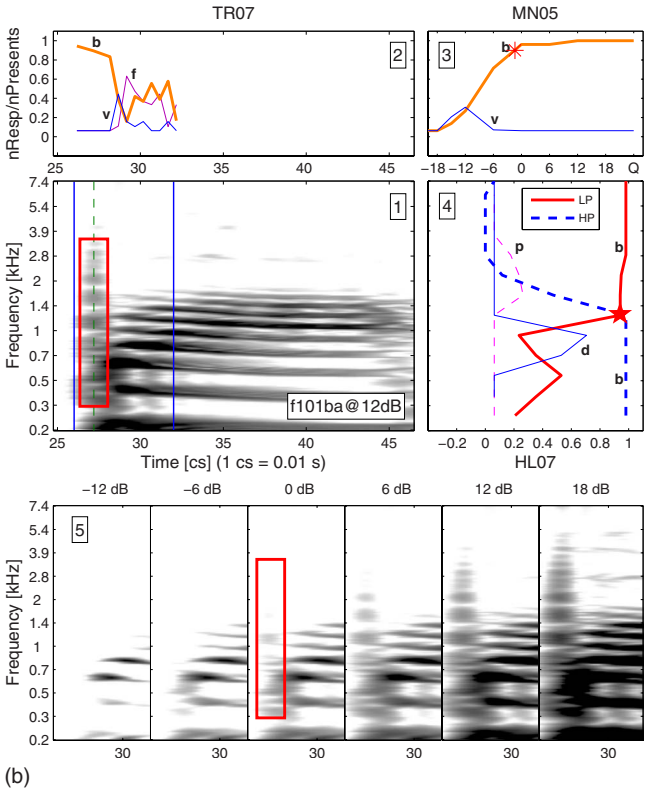


FIG. 4. (Color online) Hypothetical events for mid-frequency stop consonants /ka/ and /ga/. The multiple panels in each subfigure are 1 AI-gram. A dashed vertical line labels the onset of voicing (sonorance), indicating the start of the vowel. The solid and dashed rectangular boxes indicate the dominant and minor events, respectively. The ellipses denote the conflicting cues that cause the confusions. 2 CPs as a function of truncation time t_n. 3 CPs as a function of SNR_k. 4 CPs as a function of cutoff frequency f_k. 5 AI-grams of the consonant region (defined by the solid vertical lines on panel 1) at −12, −6, 0, 6, 12, and 18 dB SNRs.

FIG. 5. (Color online) Hypothetical events for low-frequency stop consonants /pa/ and /ba/. The multiple panels in each subfigure are 1 AI-gram. A dashed vertical line labels the onset of voicing (sonorance), indicating the start of the vowel. The solid and dashed boxes indicate the dominant and minor events, respectively. 2 CPs as a function of truncation time t_n. 3 CPs as a function of SNR_k. 4 CPs as a function of cutoff frequency f_k. 5 AI-grams of the consonant region (defined by the solid vertical lines on panel 1) at −12, −6, 0, 6, 12, and 18 dB SNRs.

same time the /ba/ → /va/ confusion $c^T_{v|b}$ for and /ba/ → /fa/ confusion $c^T_{f|b}$ rapidly increase, indicating that the wide band click is important to distinguish of /ba/ from the two fricatives /va/ and /fa/. Panel 4 shows that the highpass and lowpass scores cross each other at 1.3 kHz, the center frequency of F_2 and change abruptly, indicating that the listeners need to hear both F_2 and a significant segment of the wide band click to decide the stimulus as a /ba/. From the AI-grams in panel 5, the wide band click becomes masked by the noise somewhere below 0 dB SNR. Accordingly the listeners begin to lose /ba/ sound at the same SNR, as represented by an ∗ in panel 3. Once the wideband click has been masked, the confusions with /va/ increase and become equal to /ba/ at −12 dB SNR with a score of 40%.

There are only three LDC /ba/ sounds out of 18 with 100% scores at and above 12 dB SNR, i.e., /ba/ from f101/ shown here, and /ba/ from f109 and m120, all have a salient wide band click at the beginning. The remaining /ba/ utterances have /va/ confusions between 5% and 20%, in quiet. These nonzero quiet errors are the main difficulty in identifying the /ba/ event with certainty since the 3DDS method requires 100% in quiet for its proper operation. We do not know if it is the recordings in the LDC database that are responsible for these low scores, or if /ba/ is inherently difficult. A few high-error consonants with error rates greater than 20% were observed in LDC by Phatak and Allen (2007). From unpublished research, and not fully described here, we have found that in order to achieve a high quality /ba/ (defined as 100% identification in quiet), the wide band burst must exist over a wide frequency range. For example, a well defined 3 cs click from 0.3 to 8 kHz will give a strong percept of /ba/, which if missing or removed, may likely be heard as /va/ or /fa/.

D. Event variability

A significant characteristic of natural speech is the large variability of the acoustic cues across utterances. Typically this variability is characterized using the spectrogram. For this reason, we designed the experiment by manually selecting six utterances to have their natural variability, representative of the corpus. Since we did not, at the time, know the exact acoustic features, this was a design variable.

The center frequency of the burst (click) and the time difference between the burst and voicing onset for the 36 utterances are depicted in Fig. 6. Only the /ba/ from talker f101 is included because others do not have a wide band click and therefore highly confused with /va/ even in quiet. The figure shows that the burst times and frequencies for stop consonants are generally separated across the six different consonants.

E. Robustness

We have shown that the robustness of each consonant, as characterized by SNR_{90} is determined mainly by the strength of a single dominant cue. It is common to see the 100% recognition score drops abruptly within 6 dB, when the masking noise reaches the threshold of the dominant cue. The same observation was reported by Régnier and Allen (2008) that the 90% threshold (SNR_{90}) is directly proportional to the audible threshold of the /t/ burst based on the prediction of the AI-gram. This simple rule generalizes to the remaining five stop consonants. Figure 7 is the scatter plot of SNR_{90} versus the audibility threshold of the dominant cue. For a particular utterance (a point on the plot), the psychological threshold SNR_{90} is interpolated from the PI function), while the threshold of audibility for the dominant cue is estimated from the AI-gram. The two sets of threshold are nicely correlated over a 20 dB range, indicating that the recognition of each stop consonant is mainly dependent on the audibility of the dominant cue. Speech sounds with stronger cues are easier to hear in noise than weaker cues because it takes more noise to mask them.

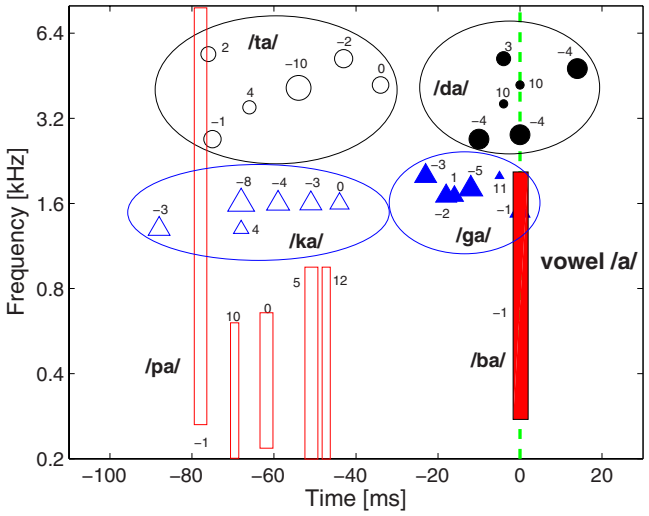


FIG. 6. (Color online) Variability of the bursts for stop consonants from multiple talkers. The strength of a burst, measured by the detection threshold (dB SNR) in white noise, is denoted by the neighboring digit.

F. Conflicting cues

It is interesting to see that many speech sounds contain conflicting cues. Take f103ka [Fig. 4(a)], for example. In

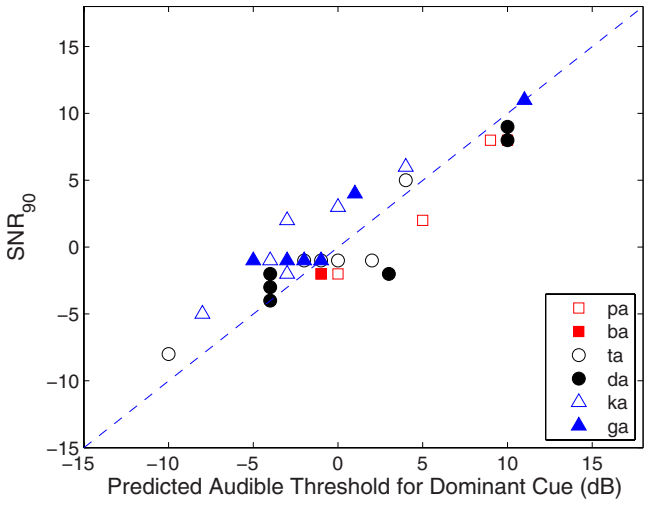


FIG. 7. (Color online) Correlation between the threshold of consonant identification and the predicted audible threshold of dominant cues based on the AI-gram.

addition to the mid-frequency /ka/ burst, it also contains two bursts in the high- and low-frequency ranges that greatly increase the probability of perceiving the sound as /ta/ and /pa/, respectively. When the dominant cue becomes masked by noise, the target sound is easily confused with other consonants within the same group. The conflicting cues have little impact on speech perception when the dominant cue is available. However, when the dominant cue is masked, the conflicting cues can cause the sound morph to another consonant. The masking range of a feature is typically 6 dB, and not more. Thus event detection is an all or nothing binary task. The spread of the event threshold is 20 dB, not the masking of a single cue. The existence of conflicting cues could make automatic speech recognition much more difficult, especially during training, because the training must sort out these false cues from the true target cues.

V. GENERAL DISCUSSION

The speech events are the perceptual information bearing aspects of the speech code. From what we have found, the density of the acoustic cues that support the events has a very low density in time-frequency space.

It was shown by Shannon (1948a, 1948b) that the performance of a communication system is dependent on the code of the symbols to be transmitted. The larger the "distance" between two symbols, the less likely the two will be confused. Shannon's proof of this principle equally applies to the case of human speech perception. For example, the /pa, ta, ka/ have common perceptual cues, i.e., a burst followed by a sonorant vowel. Once the burst is removed or masked by noise, the three sounds are highly confusable.

In all the speech perception tests, /pa, ta, ka/ commonly form a confusion group. This can be explained by the fact that the stop consonants share the same type of event patterns. The relative timing for these three unvoiced sounds is nearly the same. The major difference lies in the center frequencies of the bursts, with /pa/ having a click or low-frequency burst, /ka/ burst in the mid-frequency, and /ta/ burst in the high frequency. Similar confusions are observed for the voiced stop consonants /da/ and /ga/.

An especially interesting case is the confusions between /ba/ and /va/ [Fig. 5(b)]. Traditionally these two consonants were attributed to two different confusion groups based on their articulatory and distinctive features. However, in our experiments, we find that consonants with similar events tend to form a confusion group. Thus /ba/ and /va/ are highly confusable with each other because they share a common F_2 transition. This is strong evidence that events, not distinctive features, are the basic units for speech perception.

A. Summary

The six stop consonants are defined by a short duration burst (e.g., 2 cs), characterized by its center frequency (high, medium, and wide band), and the delay to the onset of voicing. This delay, between the burst and the onset of sonorance, is a second parameter called "voiced/unvoiced."

There is an important question about the relevance of the wide band click at the onset of the bilabial consonants /p/ and /b/. For /pa/ this click *appears* to be an option that adds salience to the sound. For /ba/ our source data are clearly insufficient because the /ba/ sounds that we chose were of poor quality; we hypothesize based on all the available data that the click is the key to a high quality /ba/ event, without which the unvoiced bilabial /ba/ is often confused with the fricatives /v/ and /f/, seen in many CPs.

In contrast, /ta/ and /ka/ are dominated by the burst frequency and delay to the sonorant onset. The voiced and unvoiced stops differ in the duration between the burst and the voicing onset. Confusion is much more common between /g/ and /d/ than with /t/ and /k/.

In other experiments, we have tried shifting the burst along the frequency axis, reliably morphing /ka/ into /ta/ (or *vice versa*). When the burst of /ka/ or /ta/ is masked or removed, the auditory system is sensitive to residual transitions in the low frequency, which cause the sound to morph to /pa/. Similarly we can convert /ga/ into /da/ (or vice versa) by using the same technique. The unvoiced stop consonants /p, t, k/ can be converted to their voiced counterpart /b, d, g/ by reducing the duration between the bursts and the onset of sonorance.

The timing, frequency, and intensity parameters may change, to a certain degree, in conversational speech, depending on the preceding and following vowels, and other factors. In a recent experiment, we investigate the effect of coarticulation on the consonant events. Instead of using vowel /a/, multiple vowels on the vertexes of the vowel triangle were selected for the study. Compared to the identified events for stops preceding vowel /a/, the identified bursts generally shift up in frequency for high vowels such as /i/ but change little for low vowels such as /u/. These recent results will be presented in a future paper.

B. Limitations of the method

It is important to point out that the AI-gram is imperfect, in that it is based on a linear model which does not account for cochlear compression, forward masking, upward masking, and other well known nonlinear phenomena seen in the auditory-nerve responses. These important nonlinearities are discussed in length in many places, e.g., Harris and Dallos (1979); Duifhuis (1980); Delgutte (1980); Allen (2008). A major extension of the AI-gram is in order, but not easily obtained. We are forced to use the linear version of the AI-gram until a fully tested time-domain nonlinear cochlear model becomes available. The model of (Zilany and Bruce, 2006) is a candidate for such testing.

Nevertheless, based on our many listening tests, we believe that the linear AI-gram generates a useful threshold approximation (Lobdell, 2006, 2008; Régnier and Allen, 2008). It is easy (trivial) to find cases where time-frequency regions in the speech signals are predicted audible by the AI-gram, but when removed, results in a signal with inaudible differences. In this sense, the AI-gram contains a great deal of "irrelevant" information. Thus it is a gross "overpredictor" of audibility. There are rare cases where the AI-gram "underpredicts" audibility, namely, where it fails to show an audible response, yet when that region is removed, the modi-

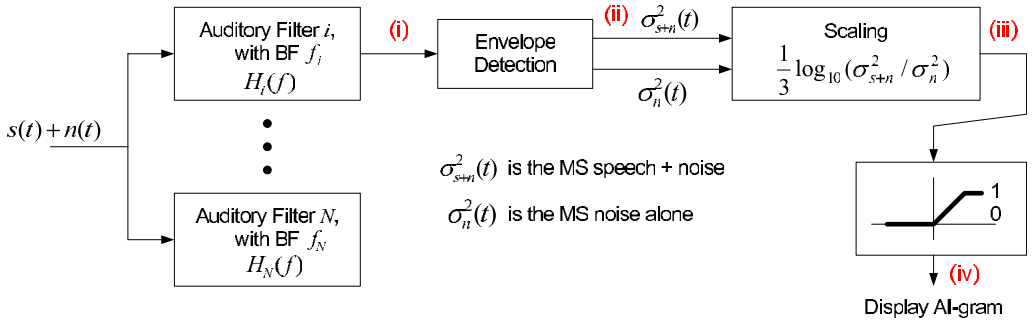


FIG. 8. (Color online) Block diagram of AI-gram [modified from (Lobdell, 2008), with permission].

fied signal is audibly different. Such cases, to our knowledge, are rare, but when discovered, are examples of serious failures of the AI-gram. This is more common below 1 kHz.

Finally, and perhaps most important, the relative strengths of cues can be misrepresented. For example, it is well know that onsets are strongly represented in neural responses due to adaptation (Delgutte, 1980). Such cues are not properly present in the AI-gram, and this weakness may be easily fixed, using existing hair-cell and neural models.

ACKNOWLEDGMENTS

The authors wish to express their appreciation to Bryce E. Lobdell, Andrea Trevino, Abhinauv Kapoor, Len Pan, Roger Serwy, and other members of the HSR group at University of Illinois, Urbana. A very small portion of this research has been supported by the NIH under Grant No. RDC009277A, awarded on 07/31/2008. This study represents a part of the Ph.D. thesis work of the first author (F.L.). They would like to acknowledge Etymotic Research and Phonak for their generous support.

APPENDIX A: THE AI MODEL

Fletcher's AI model is an objective appraisal criterion of speech audibility. The basic concept of AI is that every critical band of speech frequencies carries a contribution to the total index, which is independent of the other bands with which it is associated and that the total contribution of all bands is the sum of the contribution of the separate bands.

Based on the work of speech articulation over communication systems (Fletcher and Galt, 1950; Fletcher, 1995), French and Steinberg developed a method for the calculation of AI (French and Steinberg, 1947).

$$\mathrm{AI(SNR)} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{AI}_k, \tag{A1}$$

where AI_k is the *specific* AI for the kth articulation band (Kryter, 1962; Allen, 2005b), and

$$\mathrm{AI}_k = \min\left[\frac{1}{3}\log_{10}\left(\frac{\sigma^2_{s+n}}{\sigma^2_n}\right), 1\right], \tag{A2}$$

where $\sigma^2_{s+n}/\sigma^2_n$ is the mean-square speech+noise over noise power ratio in the kth frequency band (French and Steinberg, 1947).

Given AI(SNR) for the noisy speech, the predicted average speech error is (Allen, 1994, 2005b)

$$\hat{e}(\mathrm{AI}) = e_{\min}^{\mathrm{AI}} \cdot e_{\mathrm{chance}}, \tag{A3}$$

where $e_{\min}$ is the maximum full-band error when AI=1, and e_{chance} is the probability of error due to uniform guessing (Allen, 2005b).

APPENDIX B: THE AI-GRAM

The AI-gram is the integration of the Fletcher's AI model and a simple linear auditory model filter-bank [i.e., Fletcher's SNR model of detection (Allen, 1996)]. Figure 8 depicts the block diagram of AI-gram. Once the speech sound reaches the cochlea, it is decomposed into multiple auditory filter bands, followed by an "envelope" detector. Fletcher audibility of the narrow-band speech is predicted by the formula of specific AI [Eq. (A2)]. A time-frequency pixel of the AI-gram (a two-dimensional image) is denoted $\mathrm{AI}(t,f)$, where t and f are the time and frequency, respectively. 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 bandwidth of Fletcher (1995).

The average of the AI-gram over time and frequency, and then averaged over a phonetically balanced corpus, yields a quantity numerically close to the AI as described by Allen (2005b). An average across frequency at the output of the AI-gram yields the *instantaneous* AI

$$a(t_n) \equiv \sum_k \mathrm{AI}(t_n, f_k) \tag{B1}$$

at time t_n.

Given a speech sound, the AI-gram model provides an approximate "visual detection threshold" of the audible speech components available to the central auditory system. It is silent on which component is relevant to the speech event. To determine the relevant cues, it is necessary to directly relate the results of speech perception experiments (events) with the AI-grams (or perhaps some future nonlinear extensions of the AI-gram).

Allen, J. B. (**1977**). "Short time spectral analysis, synthesis, and modification by discrete Fourier transform," IEEE Trans. Acoust., Speech, Signal Process. **25**, 235–238.
Allen, J. B. (**1994**), "How do humans process and recognize speech"?,"

IEEE Trans. Speech Audio Process. **2**, 567–577.
Allen, J. B. (**1996**). "Harvey Fletcher's role in the creation of communication acoustics," J. Acoust. Soc. Am. **99**, 1825–1839.
Allen, J. B. (**2005a**), *Articulation and Intelligibility* (Morgan and Claypool, LaPorte, CO).
Allen, J. B. (**2005b**). "Consonant recognition and the articulation index," J. Acoust. Soc. Am. **117**, 2212–2223.
Allen, J. B. (**2008**). "Nonlinear cochlear signal processing and masking in speech perception," in *Springer Handbook on Speech Processing and Speech Communication*, edited by J. Benesty and M. Sondhi (Springer, Heidelberg, Germany), Chap. 3, pp. 1–36.
Allen, J. B., and Li, F. (**2009**). "Speech perception and cochlear signal processing," IEEE Signal Process. Mag. **29**, 117–123.
Allen, J. B., and Rabiner, L. R. (**1977**). "A unified approach to short-time Fourier analysis and synthesis," Proc. IEEE **65**, 1558–1564.
Allen, J. B., Régnier, M., Phatak, S., and Li, F. (**2009**). "Nonlinear cochlear signal processing and phoneme perception," in *Proceedings of the 10th Mechanics of Hearing Workshop*, edited by N. P. Cooper and D. T. Kemp (World Scientific, Singapore), pp. 95–107.
Blumstein, S. E., and Stevens, K. N. (**1979**). "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," J. Acoust. Soc. Am. **66**, 1001–1017.
Blumstein, S. E., and Stevens, K. N. (**1980**). "Perceptual invariance and onset spectra for stop consonants in different vowel environments," J. Acoust. Soc. Am. **67**, 648–662.
Blumstein, S. E., Stevens, K. N., and Nigro, G. N. (**1977**). "Property detectors for bursts and transitions in speech perceptions," J. Acoust. Soc. Am. **61**, 1301–1313.
Cooper, F., Delattre, P., Liberman, A., Borst, J., and Gerstman, L. (**1952**). "Some experiments on the perception of synthetic speech sounds," J. Acoust. Soc. Am. **24**, 597–606.
Delattre, P., Liberman, A., and Cooper, F. (**1955**). "Acoustic Loci and translational cues for consonants," J. Acoust. Soc. Am. **27**, 769–773.
Delgutte, B. (**1980**). "Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers," J. Acoust. Soc. Am. **68**, 843–857.
Drullman, R., Festen, J. M., and Plomp, R. (**1994a**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**, 1053–1064.
Drullman, R., Festen, J. M., and Plomp, R. (**1994b**). "Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am. **95**, 2670–2680.
Duifhuis, H. (**1980**). "Level effects in psychophysical two-tone suppression," J. Acoust. Soc. Am. **67**, 914–927.
Elliott, T. M., and Theunissen, F. E. (**2009**). "The modulation transfer function for speech intelligibility," PLOS Comput. Biol. **5**, e1000302.
Fant, G. (**1973**). *Speech Sounds and Features* (MIT, Cambridge, MA).
Fletcher, H. (**1995**), "Speech and hearing in communication," in *The ASA Edition of Speech and Hearing in Communication*, edited by J. B. Allen (Acoustical Society of America, New York), pp. A1–A34 and 1–487.
Fletcher, H., and Galt, R. (**1950**). "Perception of speech and its relation to telephony," J. Acoust. Soc. Am. **22**, 89–151.
Fousek, P., Svojanovsky, P., Grezl, F., and Hermansky, H. (**2004**), "New nonsense syllables database—Analyses and preliminary ASR experiments," in Proceedings of the International Conference on Spoken Language Processing (ICSLP).
French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.
Furui, S. (**1986**). "On the role of spectral transition for speech perception," J. Acoust. Soc. Am. **80**, 1016–1025.
Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.
Harris, D. M., and Dallos, P. (**1979**). "Forward masking of auditory nerve fiber responses," J. Neurophysiol. **42**, 1083–1107.
Hazan, V., and Rosen, S. (**1991**). "Individual variability in the perception of cues to place contrasts in initial stops," Percept. Psychophys. **59**(2), 187–200.
Heinz, J., and Stevens, K. (**1961**). "On the perception of voiceless fricative consonants," J. Acoust. Soc. Am. **33**, 589–596.
Hughes, G., and Halle, M. (**1956**). "Spectral properties of fricative consonants," J. Acoust. Soc. Am. **28**, 303–310.
Kryter, K. D. (**1962**). "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**, 1689–1697.
Liberman, A. (**1957**). "Some results of research on speech perception," J. Acoust. Soc. Am. **29**, 117–123.
Lobdell, B. E. (**2006**), "Information theoretic tool for investigating speech perception," MS thesis, University of Illinois at Urbana-Champaign, Urbana, IL.
Lobdell, B. E. (**2008**), "Information theoretic comparisons between speech intelligibility predictors and human phone perception with applications for feature extraction," Ph.D. thesis, University of Illinois at Urbana-Champaign, Urbana, IL.
Malécot, A. (**1973**). "Computer-assisted phonetic analysis techniques for large recorded corpuses of natural speech," J. Acoust. Soc. Am. **53**, 356.
Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.
Phatak, S., and Allen, J. B. (**2007**). "Consonant and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am. **121**, 2312–2326.
Phatak, S., Lovitt, A., and Allen, J. B. (**2008**). "Consonant confusions in white noise," J. Acoust. Soc. Am. **124**, 1220–1233.
Potter, R. K., Kopp, G. A., and Kopp, H. G. (**1966**). *Visible Speech* (Dover, New York).
Recasens, D. (**1983**). "Place cues for nasal consonants with special reference to Catalan," J. Acoust. Soc. Am. **73**, 1346–1353.
Régnier, M. S., and Allen, J. B. (**2008**). "A method to identify noise-robust perceptual features: Application for consonant /t/," J. Acoust. Soc. Am. **123**, 2801–2814.
Remez, R., Rubin, P., Pisoni, D., and Carrell, T. (**1981**). "Speech perception without traditional speech cues," Science **212**, 947–949.
Shannon, C. E. (**1948a**), "The mathematical theory of communication," Bell Syst. Tech. J. **27**, 379–423.
Shannon, C. E. (**1948b**), "A mathematical theory of communication," Bell Syst. Tech. J. **27**, 623–656.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Stevens, K. N., and Blumstein, S. E. (**1978**). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. **64**, 1358–1369.
Wang, M. D., and Bilger, R. C. (**1973**). "Consonant confusions in noise: A study of perceptual features," J. Acoust. Soc. Am. **54**, 1248–1266.
Zilany, M., and Bruce, I. (**2006**). "Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery," J. Acoust. Soc. Am. **120**, 1446–1466.