A method to identify noise-robust perceptual features: Application for consonant /t/

Marion S. Régnier and Jont B. Allen[a)]
ECE Department and The Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, 405 North Mathews, Urbana, Illinois 61801, USA

(Received 11 August 2007; revised 24 February 2008; accepted 25 February 2008)

This study focuses on correlating speech *confusion patterns*, defined as consonant-vowel confusion as a function of the speech-to-noise ratio, and a model acoustic feature (AF) representation called the *AI gram*, defined as the *articulation index density* in the spectrotemporal domain. By collecting many responses from many talkers and listeners, the AF and psychophysical feature (*event*) is shown to be correlated via the AI-gram model and the confusion matrices at the utterance level, thereby explaining the listener confusion. Consonant /t/ is used as an example to identify its primary robust-to-noise feature, and a precise correlation of the acoustic information with the listeners' confusions is used to label the event. The main spectrotemporal cue defining the /t/ event is an across-frequency temporal coincidence, wherein frequency spread and robustness vary across utterances, while the event remains invariant. The cross-frequency timing event is shown to be the key perceptual feature for consonants in a vowel following context. Coincidences are found to form the basic element of the auditory object. Neural circuits used for coincidence in binaural processing for localization across ears are proposed to be used within one ear across channels. It is further concluded that the event is based on the audibility of the /t/ burst rather than on any superthreshold property. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2897915]

PACS number(s): 43.71.An, 43.71.Gv, 43.66.Ba [ADP] Pages: 2801–2814

I. INTRODUCTION

After 50 years of work, even a basic understanding of speech's robustness to masking noise remains a mystery. Having a theory of human speech recognition (HSR) is critical for the development of new hearing aids and cochlear implants, and for automatic (machine) speech recognition (ASR). Given the importance of such a theory, other than a hand full of papers, little attention has been paid to the basic relations between acoustic and perceptual speech features. To research this long-standing problem, we measured a large number of listeners' responses to individual consonant-vowel (CV) syllables in noise and then correlated the confusions with each utterances' acoustic cues. *Utterances* are defined as CV sounds spoken by a particular talker, whereas a *token* is an utterance specified by its signal-to-noise ratio. We can prove the existence of these perceptual cues, denoted here as *events*, by identifying the spectrotemporal features used by listeners to discriminate consonants in noise.

Our approach is to correlate tokens to listeners's responses. By assuming that human communication is an "information channel" in the Shannon (1948) sense, we use Shannon's receiver to model the robust detection of speech cues in noise. We specifically show that across-frequency onset timing plays a crucial role in speech perception (Heil, 2003).

One might reasonably wonder why we study CV phonology rather than meaningful sounds having context. Context effects are critical when decoding natural language. However, with very limited context, listeners are still able to discriminate nonsense CV speech sounds below −16 dB speech-to-noise ratio (SNR) (Allen, 2005a; Phatak and Allen, 2007). This is now clear from our analysis of confusion matrices (CMs) of CV sounds, measured by Miller and Nicely (1955) (denoted MN55), as emphasized by Allen (2005a, 2005b) and Phatak and Allen (2007). Noise robustness has been a major area of debate in both the HSR and ASR literature (Hermansky, 1998).

Other than in a handful of studies (Blumstein and Stevens, 1979; Delattre *et al.*, 1955; Repp *et al.*, 1978; Kewley-Port, 1983; Lisker, 1985; Stevens and Klatt, 1974; Klunder *et al.*, 1995) little is known about the specific spectrotemporal information present in each wave form that causes (or prevents) specific confusions. Much remains to be done. Only a few studies (Fletcher and Galt, 1950; Summerfield and Haggard, 1977; Dubno and Levitt, 1981; Dubno *et al.*, 1987; Kamm *et al.*, 1985; Hant and Alwan, 2003) have used masking noise on real speech signals to study the confusions via the confusion matrix method. To explore the effects of noise and its spectrum, the University of Illinois at Urbana-Champaign (UIUC) Human Speech Recognition Group at the Beckman Institute conducted two basic studies, denoted PA07 and PA05 (Lovitt and Allen, 2006) and analyzed by Phatak and Allen (2007). PA05 used the same CVs as MN55 presented in white noise but expanded the MN55 experiment by including more vowels, talkers, and listeners and by updating the testing methodology. The present paper extends the analysis of Phatak and Allen (2007) by correlating the audible speech information with the scores for individual utterances from these two experiments. Our goal is to identify and label the common robust-to-noise features in the spectrotemporal domain (Strope and Alwan, 1997).

[a)]Electronic mail: jontalle@uiuc.edu.

Previous studies [Cooper *et al.*, 1952; Delattre *et al.*, 1955; see Hawkins, 2003; Nguyen and Hawkin, 2003 for reviews] pioneered the analysis of spectrotemporal cues discriminating consonants. Their goal was to study the acoustic properties of consonants /p/, /t/, and /k/ in different vowel contexts. One of their main results is the empirical establishment of a physical to perceptual map, derived from the presentation of synthetic CVs to human listeners. These synthetic "speech sounds" consisted of a short noise burst (10 ms, 400 Hz bandwidth), representing the consonant, followed by artificial formant transitions composed of tones, simulating the vowel. They discovered that for each of these synthetic voiceless stops, the spectral position of the noise burst was vowel dependent. This *coarticulation* was mostly obvious for /p/ and /k/, with bursts above 3 kHz giving the percept of /t/ for all vowel contexts. A burst located at the second formant frequency or slightly above would create a percept of /k/, and that below would create /p/. Consonant /t/ could therefore be considered to have less coarticulation. No information was provided about the robustness of their synthetic speech samples to masking noise nor about the importance of the presumed features relative to other cues that are present in natural speech but missing in their synthetic speech samples.

The spectrotemporal location of events has been found to vary due to the natural variability of speech. Cooper *et al.* (1952) determined the most relevant parts of the speech based on perceptual criterion. We have done the same. However, unlike Cooper *et al.* (1952), our results depend on natural speech and variable amounts of masking noise (Miller and Nicely, 1955; Allen, 2005a; Phatak and Allen, 2007; Régnier and Allen, 2007a, 2007b).

In summary, by collecting many responses from many talkers and listeners, we were able to build a large database of CPs (Allen, 2005b; Phatak and Allen, 2007). We relate the perceptual and physical domains at the utterance level via our measurement of speech audibility, the articulation index (AI) model (Fletcher and Galt, 1950; French and Steinberg, 1947), and thereby explain listeners' confusions via the spectrotemporal acoustic perceptual features. Throughout this paper, we will take the example of consonant /t/ and show how we can reliably identify its primary robust-to-noise feature. In order to identify and label events, we precisely correlate the acoustic information with the listeners' confusions. We show that the main spectrotemporal cue defining the /t/ event is composed of across-frequency temporal coincidence (perceptual features), which is represented by correlated acoustic properties (acoustic features). These can vary on an utterance basis. Our observations support these coincidences as a basic element of the auditory object formation, the cross-frequency timing event being the main perceptual feature for consonants in a vowel context. It seems reasonable that similar neural circuits used for coincidence in binaural processing for localization across ears (Joris *et al.*, 1998) could be used within one ear across channels (Delgutte *et al.*, 1998).

II. THE ARTICULATION INDEX: AN AUDIBILITY MODEL

The *articulation* is defined as the score P_c(SNR) for nonsense sound. The *articulation index* is the foundation stone of speech perception, laid down by Fletcher and Galt (1950) and French and Steinberg (1947). The AI is a *sufficient statistics* of the articulation (Allen, 2005a). It follows that the score is a function of the AI, giving P_c(AI). The basic concept of the AI is to quantify maximum entropy (MaxEnt, also called nonsense speech) average phone scores. It is based on the SNR in critical bands, expressed in decibel sensation level, scaled by the dynamic range of speech (30 dB) (Allen, 1994, 2005a; Phatak and Allen, 2007).

Allen (2005a) showed that French and Steinberg's (1947) expression for the maximum entropy average phone score P_c(AI), corrected for chance, may be written as

$$P_c(\mathrm{AI}) = 1 - P_e = 1 - e_{\mathrm{chance}} e_{\mathrm{min}}^{\mathrm{AI}}, \tag{1}$$

where the recognition error e_{min} is the minimum error at AI=1, and the error $e_{\mathrm{chance}}=1-1/16=15/16$ at chance performance (AI=0) for the 16 consonant case (Kamm *et al.*, 1985).

The articulation index is the basis of many standards, and its long history and utility, as discussed in length in several papers and books (French and Stemberg, 1947; Allen, 1994; Fletcher, 1995; Allen, 1996, 2005a, 2000b). Phatak and Allen (2007) extended the AI formula following French and Steinberg (1947) to account for the peak-to-rms [root-mean-squared $x_{\mathrm{rms}}=\sqrt{\overline{(x-\bar{x})^2}}$] ratio for the speech r_k in each band as

$$\mathrm{AI}_k = \min\left(\tfrac{1}{3}\log_{10}(1 + r_k^2 \mathrm{SNR}_k^2), 1\right), \tag{2}$$

where AI_k is called the *specific AI*, and SNR_k is the SNR (i.e., the ratio of the rms of the speech to the rms of the noise) in the kth articulation band. The total AI is then the average over the specific AI:

$$\mathrm{AI} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{AI}_k. \tag{3}$$

The parameter $K=20$ bands defines the number of articulation bands and is determined empirically to give an equal band contribution in score for consonant-vowel materials (French and Steinberg, 1947). Each of these 20 bands corresponds to 1 mm along the basilar membrane, from about 0.3 to 7.5 kHz (Fletcher and Galt, 1950).

We denote the AI density over time and place (frequency) as the *AI gram* and noted it as $\mathrm{AI}(t, X_k)$. This is a simple graphical extension of the AI, as defined by French and Steinberg (1947). This function of time and *place* (defined as the distance X along the basilar membrane) is computed from a cochlear model or cochlear filter bank, with bandwidths equal to human critical bands, followed by a simple model of the auditory nerve (Lobdell and Allen, 2006). Figure 1 shows the block diagram of how the AI gram is computed from a masked speech signal $s(t)$. As a practical matter, the noise spectrum $\sigma_n^2(f)$ is based on the noise in isolation, but if necessary, it may be estimated directly from the noisy speech but with some loss of accuracy. The AI gram includes a conversion of the basilar membrane vibration to a neural firing rate via an envelope detector representing the mean rate of the neural firing pattern across the cochlear output. The speech+noise signal is scaled by the long-

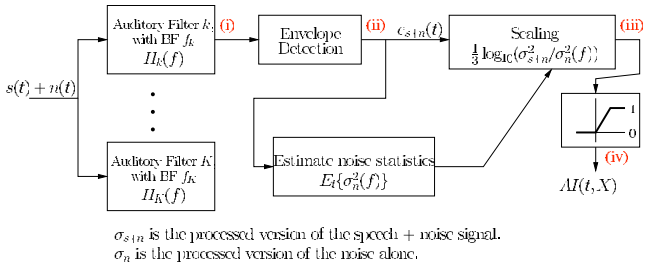

FIG. 1. (Color online) AI-gram block diagram. Derived from the work of French and Steinberg (1947) and Fletcher (1995), the output of the AI gram relates to the audibility of a sound. When a speech signal is visible in different degrees of black on the AI gram, it is above the masked threshold (i.e., audible). It follows that all noise and inaudible sounds having a SNR less than 0 dB appear in white due to the band normalization by the noise.

term average noise power $\sigma_n^2(f)$ so that the power SNR in each frequency band is $\mathrm{SNR}_k^2=\sigma_{s+n}^2/\sigma_n^2\approx 1+\sigma_s^2/\sigma_n^2$. The scaled value of SNR_k (in decibels) [as defined by Eq. (2)] yields the AI density $\mathrm{AI}(t,X)$. The audible speech modulations across frequency give a spectrotemporal representation in the form of the AI gram (Lobdell and Allen, 2006), as shown in Fig. 2.

The AI gram represents a simple *perceptual model*, and based on the pioneering work of French and Steinberg (1947) and Fletcher and Galt (1950), its output is assumed to be correlated with our psychophysical experiments. Different degrees of black on the AI gram are correlated with the audibility of the region of interest. Noise and inaudible speech sounds are represented by white (where the SNR is less than 0 dB), while audible speech is dark. The greater the $\mathrm{AI}(t,X)$, the darker the corresponding region.

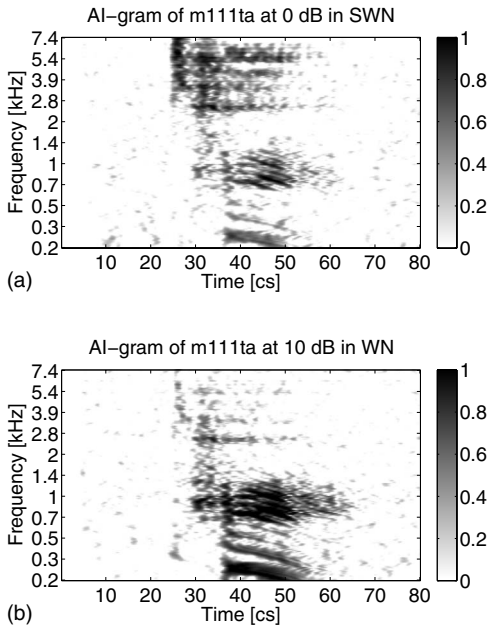

FIG. 2. AI gram of male speaker 111 speaking /tɑ/ in (a) SWN at 0 dB SNR and (b) WN at 10 dB SNR. The audible speech information is dark, the different levels representing the degrees of audibility. Since the two different noises have different spectra, the speech is masked differently. Speech-weighted noise masks low frequencies and high frequencies equally on average. One may clearly see the strong masking due to WN at high frequencies. The AI gram is an important tool used here to explain the differences in CPs observed in many studies to connect the physical and perceptual domains.

Figure 2 provides two AI grams of utterance /tɑ/ in (a) speech-weighted noise (SWN) and (b) white noise (WN). We shall demonstrate that this model of audibility is useful when explaining the decrease of recognition and for identifying the correlated masked events.

III. PILOT EXPERIMENTS

The purpose of this pilot study section is to draw out previously unpublished results from two major studies, called PA05 and PA07 here. PA05 is a modern version of the study of Miller and Nicely (1955). Results are unpublished. It uses 16 consonants and a single vowel with white noise. PA07 is an extension of MN55, with 64 CVs (16 consonants and 4 vowels), using speech-weighted noise (Phatak and Allen, 2007). Both experiments are further analyzed here in terms of confusion patterns (CPs) $P_{h|s}(\mathrm{SNR})$ (Allen, 2005b) and then correlated against AI grams. We have carried out this analysis on the consonant /t/ using an analysis tool denoted the *four-step method*. This novel detailed analysis will provide some useful insights that we believe to be new.

The next section describes the methods and results of these Miller-Nicely-type experiments, PA07 and PA05, run by members of the Human Speech Recognition Group in 2004 (Phatak and Allen, 2007).

A. PA07 and PA05

Experiment PA07 measured normal hearing listeners' responses to 64 CV sounds (16 C × 4 V, spoken by 18 talkers). PA05 included the subset of these CVs containing vowel /ɑ/, to match MN55's CVs.

B. Methods

For PA07, the masking noise was SWN having SNRs of $[Q,12,-2,-10,-16,-20,-22]$ (Q for quiet), and that for PA05 WN was used with SNRs of $[Q,12,6,0,-6,-12,-15,-18,-21]$. All conditions were randomized and presented only once to each of the approximately 20 listeners. The CV speech stimuli for all experiments described in this paper are taken from the Linguistic Data Consortium at the University of Pennsylvania (corpus LDC-2005S22) (Fousek, *et al.* 2004) database composed of a large number of CV, VC, CCV, and VVC utterances, spoken by 20 talkers having different language backgrounds. Talkers' labels start with a talker gender label (f, m), followed by the talker ID (a three-digit number) and the sound label (e.g., m115pa).

The experiments were implemented using MATLAB©. The presentation program was written by student Andrew Lovitt, run on a desktop PC with a Linux kernel 2.4 (Mandrake 9) located outside an acoustic booth (Acoustic Systems model number 27930). Only the keyboard, liquid crystal display monitor, headphones, and mouse were inside the booth. Subjects sitting in the booth are presented with the speech files via a 16 bit PC sound card (Soundblaster Live) and Sennheiser HD280 headphones. Subjects responded as to which CV they heard via a graphical interface.

1. SNR calculations

To set the rms level, the following procedure was used. The rms level for the speech and the noise was set to 1 by computing the rms level for each, and dividing each waveform by that number. Second, the SNR was converted to a decimal value using the formula $g=10^{\text{snr}/20}$, where SNR is the decibel value. Third, the speech and noise were added according to the formula $\text{fix}(S=s+g*n)$ where s,n are the vector speech and noise wave forms. Finally, the absolute peak was found and the vector signal S was converted to an integer wave form, in preparation for outputting to the codec. This transformation is $KS(t)/S_{\max}$ where $K=32\,767=2^{15}-1$ is the maximum positive integer supported by the codec corresponding to the maximum codec voltage.

The rms for the speech was based on either a Volume Unit (VU) calculation (Lobdell and Allen, 2007) or a true rms calculation depending on the data set in question (Phatak and Allen, 2007). The rms of the noise was based on a true rms calculation.

To prevent loud sounds, the maximum allowed rms pressure was limited to 80 dB sound pressure level (SPL) by a hardware attenuator box located between the sound card and the headphones. This limit was based on a 1 kHz long-duration pure tone. The earphone sensitivity calibration (in Pa/V) was based on a 1 kHz tone SPL level on a flat plate coupler using an Etymotic ER7C probe microphone. When frequently asked, the subjects never requested any further level adjustment.

Subjects were volunteers from the University of Illinois student and staff population who had normal hearing (self-reported) and were native English speakers. They were compensated for their participation. Further details on methods may be found in the references.

C. Confusion pattern analysis

CPs [a row of the CM versus SNR, $P_{h|s}(\text{SNR})$] corresponding to a specific CV "spoken" (s) versus "heard" (h) utterance provide the representation of the scores as a function of SNR. CPs may be made for individual utterances or averaged over all talkers/utterances of the same CV. In the following analysis, we shall look at CPs for specific utterances.

Examples of CP plots for an individual /tɑ/ utterance from PA05 and PA07 are shown in Fig. 3, averaging data for 14 listeners for PA07 and 24 for PA05. Many important observations are supported by these charts.

1. Utterance variability

As SNR is reduced (noise increased), the target consonant score starts to measurably decrease at the *saturation threshold*, denoted SNR_s. This *robustness threshold* is defined as the SNR at which the consonant had an error equal to chance performance (6.25%). For example, in Fig. 3, corresponding to utterance f105ta, SNR_s is located at −16 dB SNR for SWN [left panel (a)] and at 0 dB SNR in WN [right panel (b)]. Each utterance presents a different threshold depending on its physical properties. This dependence will be explored in detail in the next section, Sec. III D, and then discussed in Sec. III E.

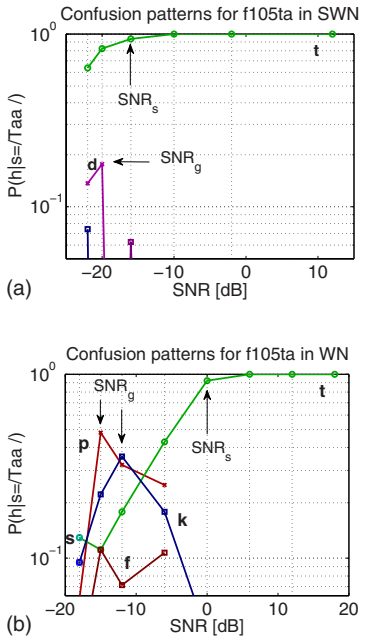

FIG. 3. (Color online) Confusion patterns for /tɑ/ spoken by female talker 105 in (a) SWN and (b) white noise. Note the significant robustness difference depending on the noise spectrum. In SWN, /t/ is correctly identified down to −16 dB SNR, whereas it starts to decrease at 0 dB in WN. SNR_s corresponds to the SNR at which the error starts to increase. The confusions are also more significant in WN, with the scores for /p/ and /k/ overcoming that of /t/ below −6 dB. We call this surprising observation morphing. The maximum confusion score is denoted SNR_g. The reasons for this robustness difference depend on the audibility of the /t/ event, which will be analyzed in the next section.

2. Spectral effects

It is clear from Fig. 3 that the noise spectrum (WN or SWN) strongly influences both SNR_s and the confusions occurring well below the saturation threshold. The *confusion group* of this /tɑ/ utterance in WN [Fig. 3(b)] is [/p/, /t/, /k/]. The SNR corresponding to the maximum in the confusion score is denoted SNR_g, and for this example is located close to (indicated with arrows) −18 dB SNR for /p/ and −15 dB for /k/, with respective scores of 50% and 35%. In SWN (a), /d/ is the only significant competitor (at a 20% level. $\text{SNR}_s=-16$ dB, $\text{SNR}_g=-20$ dB) due to its natural robustness to SWN. Summarizing then, the same utterance presents different robustness and confusion thresholds depending on the masking noise and the consonant spectral support. We shall further explore this in Sec. III D.

3. Morphing

As WN is mixed with utterance f105ta, /t/ *morphs* to /p/ or /k/, meaning that the probability of recognizing /t/ drops while that of /p/ and /k/ increases above that of /t/. At a SNR of −9 dB (WN), the /p/ and /k/ confusions overcome the target /t/ score. As shown in the CP plot of Fig. 3(b), the recognition of /p/ is maximum $[P_{/p/|/t/}(\text{SNR}_g=-16)=50\%]$, that of /k/ peaks at 35% at −12 dB, whereas the score for /t/ is about 10%. This morphing effect, while at first surprising,

is typical in our database for many consonant in white noise but much less common in SWN. The robustness of /t/ to SWN makes morphing rare in PA07 (SWN), as will be exampled later on, due the large difference in SNR spectral balance.

4. Priming

Listening experiments show that when the scores for consonants of a confusion group are similar, listeners can *prime* between these phones. Priming is defined as the ability to mentally select the consonant heard by making a conscious choice between several possibilities having neighboring scores. As a result of priming, given random presentations, a listener must randomly choose one of several priming consonants. The SNR range for which priming takes place is listener dependent; the CPs presented here are averaged across listeners and, therefore, are representative of an average priming range. At a given SNR, each listener has a bias toward one or the other sound, causing score differences. However, most listeners prime between /t/, /p/, and /k/ at around −10 dB SNR, whereas they typically have a bias for /p/ at −16 dB SNR and for /t/ above −5 dB. Based on our studies, we suspect that priming occurs when events, shared by consonants of a confusion group, are at the threshold of audibility, and when the distinguishing feature is at its masked threshold.

In summary, four major observations may be drawn from our analysis of many CPs of CVs similar to those of Fig. 3: (i) *robustness variability across utterances* as measured by SNR_s, (ii) *confusion group variability across noise spectra and utterances*, (iii) *morphing*, and (iv) *priming*. We conclude that each utterance presents different *saturation thresholds*, different *confusion groups, morphs* (or not), and may be subject to *priming* in some SNR range, depending on the masking noise spectrum and the consonant.

Our approach is to take advantage of this natural variability across tokens. As exampled in the above discussion, we will quantitatively relate the confusion patterns and robustness to the audible cues at a given SNR. Finding such relations enables us to identify events, namely, to label the acoustic features that map to "perceptual space."

Using the four-step method, described in the next section, we demonstrate that events are common across utterances of a particular consonant, whereas the *acoustic correlation* of the events, namely, the spectrotemporal and energetic properties, depends on the utterances, the noise spectrum, and the SNR. That is, while the acoustic features are highly variable, events are not. While in this report we only study /t/ and its confusions, we believe that /t/ is representative of other sounds in this regard.

5. Conclusion of confusion pattern analysis

We have used an AI-gram analysis on a large number of responses to many CV sounds in noise and related the scores to the audible speech features, with the end goal of finding events. These events represent the acoustic features that are robust to noise since they survive at very low SNRs (typically much less than 0 dB). Several features of the CPs have been defined, such as morphing, priming, and utterance variability in noise. The identification of a saturation threshold SNR_s (Fig. 3), located at the 93.75% point, is a quantitative measure of utterance robustness, and is a function of the noise spectrum. The natural utterance variability, quantified by this robustness threshold, causes utterances of the same phone category to behave differently with different types of noise.

The existence of morphing demonstrates that noise can mask an essential feature for the recognition of a sound, leading to consistent confusions among our subjects. However, such morphing is not ubiquitous, as it depends on the type and amount of masking noise. Different morphs are observed for different noise spectra. Morphing demonstrates that consonants are not uniquely characterized by independent features, but that they share *common cues* that appear to be weighted differently in perceptual space. This conclusion is supported by CP plots for /k/ and /p/ utterances, showing a well defined [/p/, /t/, /k/] confusion group structure in WN.

From the strong confusions, it seems clear that [/t/, /p/, /k/] share common perceptual features. A similar conclusion can (arguably) be derived from the results observed by Miller and Nicely (1955) (Allen, 2005b). The /t/ event is more easily masked by WN than by SWN, and the usual [/k/, /p/] confusion for /t/ in WN demonstrates that when the /t/ burst is masked, the remaining features are shared by all three voiceless stop consonants. As exampled in the next section, in Fig. 4(a), when the primary /t/ event is masked at high SNRs in SWN, we see strong [/p/, /t/, /k/] confusion groups. Our working hypothesis is that the common features shared by this group are masked by speech-weighted noise due to their localization in frequency, whereas the /t/ burst itself is usually robust in SWN.

We shall further show in Sec. III that this *common feature* hypothesis is also supported by temporal truncation experiments, similar to those of Furui (1986). It is shown that confusions take place when the acoustic features defining the primary /t/ event are inaudible due to noise or truncation, and that the remaining cues are part of what perceptually characterizes /t/'s competitors (/p/, /k/).

D. Four-step method to identify events

Our four-step method (steps 1–4) is an analysis that uses the perceptual models and correlates them with the CPs. This leads to the development of the *event gram*, derived from the AI gram, and uses human confusion responses to identify the relevant acoustic features. Here, we used the four-step method to draw conclusions about the /t/ event. There is some evidence that we may be able to explain other stop consonants using the four-step method. This will be studied in the near future. In Fig. 4(a), we identify and analyze the spectral support of the primary /t/ perceptual feature for two /tɛ/ utterances in SWN spoken by different talkers.

1. Step 1: CPs and robustness

Step 1 of our four-step analysis consists of the collection of confusion patterns described in the previous section, as in the bottom right panels of Figs. 4(a) and 4(b).

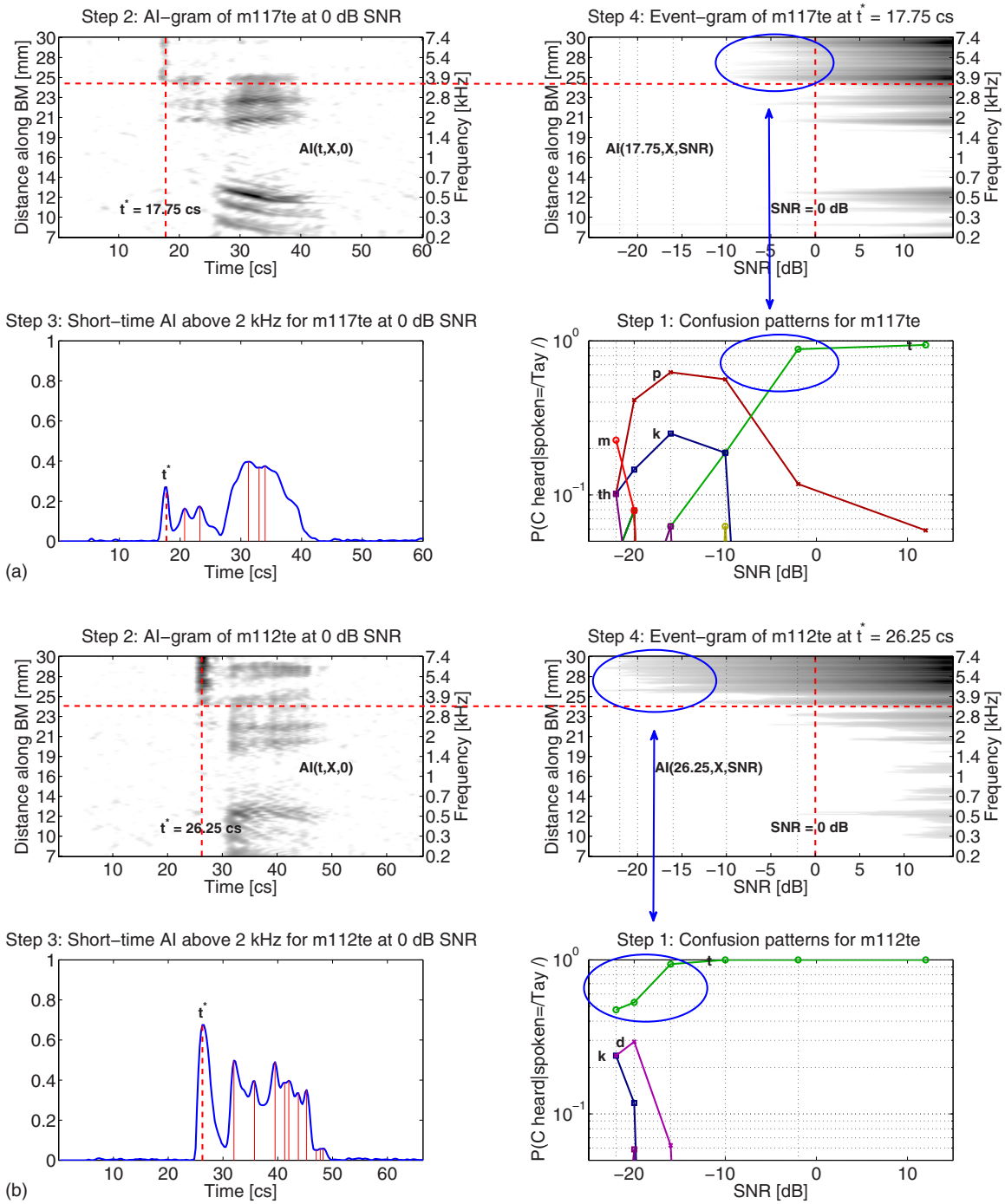

FIG. 4. (Color online) Comparison between a "weak" (top, m117te) and a "strong" (robust) (bottom, m112te) /tɛ/. The arrangement of the four panes is optimized for inner subfigure comparisons. Step 1 provides to the CPs (bottom right), step 2 to the AI gram at 0 dB SNR in SWN, step 3 to the mean AI above 2 kHz where the local maximum t^* in the burst is identified, leading to step 4, the event gram (vertical slice through AI grams at t^*). Note that for the same SWN masking noise, these utterances behave differently and present different competitors. Utterance m117te strongly morphs to /pɛ/. Many of these differences can be explained by the AI gram and more specifically by the event gram which quantifies the /t/-burst threshold, and therefore its robustness to noise. This threshold is precisely correlated with the human responses (encircled). This leads to the conclusion that this 2–8 kHz across-frequency onset transient is the primary /t/ event. (a) Analysis of sound /tɛ/ spoken by male talker 117 in SWN. This utterance is not robust to noise since the /t/ recognition starts to decrease at −2 dB SNR. Identifying t^*, time of the burst maximum at 0 dB SNR in the AI gram (top left), and its mean in the 2–8 kHz range (bottom left), leads to the event gram (top right). The vertical dashed line on the AI gram shows t^*. On the event gram, the dashed line shows the SNR at which the AI gram was displaced (similar to a vertical slice). In both cases, the horizontal dashed line marks the lower frequency limit of the burst (here 3.7 kHz) This representation of the audible phone /t/ burst information at time t^* is highly correlated with SNR_s and the CPs: when the burst information becomes inaudible (white on the AI gram), /t/ score decreases, as indicated by the ellipses. (b) Analysis of sound /tɛ/ spoken by male talker 112 in SWN. Unlike the case of m117te, this utterance is robust to SWN and identified down to −16 dB SNR. Again, the burst threshold defined by the event gram (top right) is related to SNR_s defined by the CP, accounting for the robustness of consonant /t/.

For male talker 117 speaking /tɛ/ [Fig. 4(a), bottom right panel], the saturation threshold is ≈−6 dB SNR (ellipsed), forming the /p/, /t/, /k/ confusion group. This weak /t/ morphs to /p/ [Fig. 4(a)]: the recognition of /p/ is maximum $[P_{/p/|/t/}(\mathrm{SNR}_g)=60\%]$ at a $\mathrm{SNR}_g=-16$ dB, where the score for /t/ is 6%. Morphing not only occurs in WN (Fig. 3

f105ta) but also in SWN for this weaker /tɛ/ sound (m117te). Confusion patterns and robustness vary dramatically across utterances of a given CV masked by the same noise: unlike for talker m117, /te/ spoken by talker m112 does not morph to /p/ or /k/, and its score is higher [Fig. 4(b), bottom right panel]. For this more robust utterance, /t/ was accurately identified down to $SNR_s = -16$ dB SNR (ellipsed) and was still well above chance performance (1/16) at −22 dB. Its main competitors /d/ and /k/ have 25%–30% scores and which only appear well below SNR_s.

It is clear that these two /tɛ/ sounds are dramatically different. Note that utterance differences are only seen with the addition of a masking noise. There is confusion pattern variability not only across noise spectra but also within a masking noise category (e.g., WN versus SWN). These two /tɛ/'s are an example of *utterance variability*, as shown by the analysis of step 1: two sounds are heard the same when it is quiet, but they are heard quite differently as the SNR is varied. The next section will detail the physical properties of consonant /t/ in order to relate spectrotemporal features to the score using our audibility model.

2. Steps 2 and 3: Utilization of a perceptual model

For talker 117, Fig. 4(a) (top left panel) shows the AI gram at 0 dB SNR. We observe that the high-frequency burst having a sharp energy onset, stretches from 2.8 to 7.4 kHz and runs in time from 16 to 18 cs (a duration of 20 ms). According to the CPs previously discussed [Fig. 4(a), bottom right panel], at 0 dB SNR consonant, /t/ is recognized 88% of the time. The burst for talker 112 has higher intensity and spreads from 3 kHz up, as shown by the AI gram for this utterance [Fig. 4(b), top left panel], which results in a 100% recognition at and above −10 dB SNR.

These observations lead us to step 3, the integration of the AI gram over frequency [bottom right panels of Figs. 4(a) and 4(b)]. One obtains a representation of the average audible speech information over a particular frequency range Δf as a function of time, denoted the short-time AI, AI(t) (Rhebergen *et al.*, 2006). Here, AI(t) is computed in the 2–8 kHz bands, corresponding to the limits of this high-frequency /t/ burst. In contrast, the traditional AI is the area under the entire frequency range. (In this case, if we integrated over all frequencies, there would be only a very small difference.) The first maximum, AI(t^*) [vertical dashed line on the top and bottom left panels of Figs. 4(a) and 4(b)] is an indicator of the audibility of the consonant /t/ burst. Since the frequency content is collapsed, t^* indicates the time of the relevant perceptual information for /t/.

3. Step 4: The event gram

The identification of t^* allows step 4 of our correlation analysis. The top right panels of Figs. 4(a) and 4(b) represent the event grams for the two utterances. The event gram, AI(t^*, X, SNR), is defined as a cochlear place (or frequency, via Greenwood's cochlear map) versus SNR slice at one instant of time. The event gram is the link between the CPs and the AI gram. The event gram represents the AI density as a function of SNR at a given time t^* (determined in step 3). If several AI grams computed at different SNRs were stacked, the event gram would be a vertical slice through this stack. In summary, the event grams displayed in the top right panels of Figs. 4(a) and 4(b), plotted at t^*, characterize the /t/ burst's threshold. A horizontal dashed line, from the bottom of the burst on the AI gram to the bottom of the burst on the event gram at SNR=0 dB, provides the visual link between the two plots.

The significant result visible on the event gram is that for these two utterances, the event gram is correlated with the average normal listener score, as seen in the ellipses linked by a double arrow. Indeed, for utterance 117te, the recognition of consonant /t/ starts to drop at −2 dB SNR, when the burst above 3 kHz is completely masked by the noise [top right panel of Fig. 4(a)]. On the event gram, below −2 dB SNR (circle), one can note that the energy of the burst at t^* decreases, and the burst becomes inaudible (white). A similar relation is seen for utterance 112, but since the energy of the burst is much higher, the /t/ recognition starts to fall at −15 dB SNR, at which point the energy above 3 kHz becomes sparse and decreases, as seen in the top right panel of Fig. 4(b) and highlighted by the circles.

There is an obvious correlation in this example between the variable /t/ confusions and the score for /t/ [step 1, bottom right panel of Figs. 4(a) and 4(b)], the strength of the /t/ burst in the AI gram (step 2, top left panels), and the short-time AI value (step 3, bottom left panels), all quantifying the event gram (step 4, top right panels). Because these panels are correlated with the human score, the burst constitutes our model of the perceptual cue (the event), upon which listeners rely to identify consonant /t/ in noise.

A systematic quantification of this correlation for a large number of consonants will be described in the next section, where we analyze the effect of the noise spectrum on the perceptual relevance of the /t/ burst in noise to account for the differences previously observed across noise spectra.

4. Effect of the noise samples

A classical question is the difference between internal and external noises in masking (Allen and Neely, 1997). Internal noise is due to the internal representation of the signal in a critical band while external noise is due to the uncertainty from the external noise added to the signal. Given this classical classification scheme, it was natural to wonder about the significance of the different ensemble noise samples on the variability of the event gram. Our experiment procedure was designed so that a new noise sample was used for each token presentation, so listeners never hear a signal (speech sound) mixed with the same masker (noise) even when presented at the same SNR.

To study this effect, we analyzed the variance using different noise samples having the same spectrum (the phase of the noise varied from trial to trial) and computed event grams for ten different noise samples. The resulting variance is shown in Fig. 5 for utterance f103ta in SWN. The result is that both regions of high and low SNRs show a small variance. Only where the SNR is near zero do the speech and noise interact, giving a finite variance. It makes sense that the only noticeable variance is seen near threshold. The

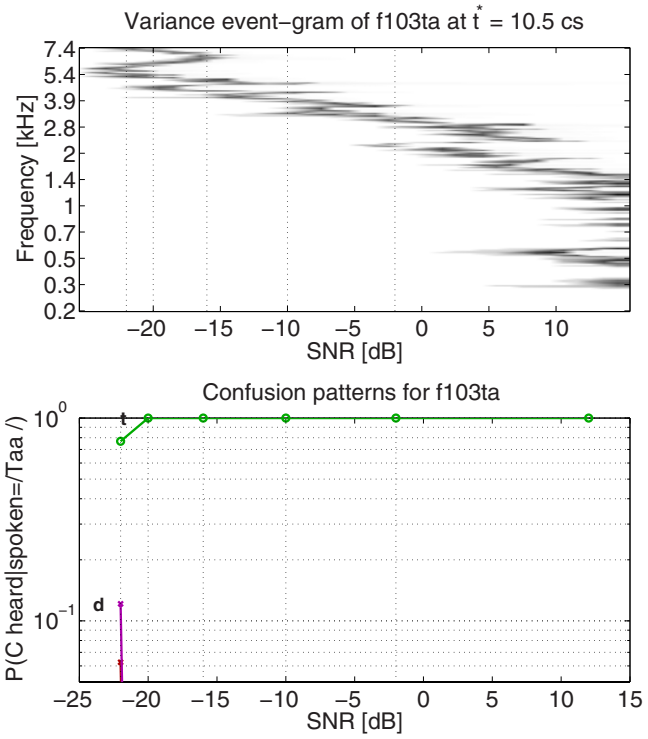

FIG. 5. (Color online) This variance event gram was computed by taking event grams of a /tɑ/ utterance for ten different noise samples in SWN (PA07). We can see that all the variance is located on the edges of the audible speech energy, where the noise and speech have similar level, located between regions of high audibility and regions of noise. However, the spread is thin, showing that the use of different noise samples will not significantly impact the perceptual scores.

thickness of the line is a measure of the trial variance. The small spread of the line indicates that using a new noise on every trial has little impact on the scores of our psychophysical experiment, and the correlation between noise and speech is unlikely to impact the features significantly, and to the extent it does, only over a small SNR range.

E. Discussion: Relating CPs and audibility for /t/

Using a four-step method analysis, we found that the discrimination of /t/ is due to its robustness, defined by a sharp onset burst from 2 to 8 kHz. Robustness, measured by SNR_s and CPs (confusion groups), are highly dependent on the specific utterance due to variations in this burst. Each instance of the /t/ burst typically presents different characteristics. As shown in Fig. 3, SNR_s changes with the noise type for a given sound. This correlation demonstrates that a consonant presents different perceptual thresholds. Additionally, in Fig. 4, we found that the /t/ event is invariant for two utterances and is masked at different SNRs and correlated with SNR_s.

As compared to SWN, WN provides more masking at high frequencies, accounting for the decrease of the /t/ at high SNR recognition. Once the burst starts being masked, at SNR_s, the /t/ score quickly drops below 100%. The acoustic representations in the physical domain of the perceptual features is highly variable (e.g., 16 dB in our example), while the perceptual features themselves (events) remain invariant.

We wish to more precisely quantify the /t/ event in the physical domain; however, improved characterization of the spectrotemporal location of the burst is needed to fully quantify its impact on robustness across utterances.

To further quantify the correlation between the audible speech information from the event gram and the perceptual information given by our listeners, we have correlated *event-gram thresholds*, denoted SNR_e, with the 90% score SNR, denoted SNR_{90}. Based on our regressions between SNR_e and the score, the 90% score point represents the /t/ feature with a smaller residual error over the wide range of conditions we studied. Thus, the onset drop in score appears to be more highly correlated with SNR_e than SNR_s. Here, we wish to identify the optimal set of parameters that will maximize the correlation between SNR_e and SNR_{90}). Our previous results, based on the four-step method, enable us to limit the search above 2 kHz, a frequency ranged *assumed* to be most relevant to /t/ recognition.

The parameters used here are an AI density threshold T and an adaptive bandwidth B. For given values of T,B, we determined the SNR range where there is continuous speech information in frequency above AI threshold T. To start the search for the optimum values of T and B, a set of values of SNR_e is determined for different pairs of parameters, within 5 Hz steps for the bandwidth and steps of 0.005 for the threshold. The value corresponding to the lowest mean square error for the correlation with SNR_{90} gives SNR_e. The procedure was followed independently for PA05 and PA07 and gave two slightly different optimized parameters, namely, $B=570$ Hz in SWN for $T=0.335$, and $B=450$ Hz for $T=0.125$ in WN.

Fourteen of the /ɑ/ utterances tested in PA07 were also in PA05; therefore, sound common to both experiments appears twice on the scatter plot. Scatter plots for PA05 (in WN) are at higher SNRs than those for PA07 (in SWN) due to the stronger masking of the /t/ burst in WN, leading to higher SNR_e and SNR_{90}.

The high correlation between SNR_{90} and the event-gram thresholds for the parameters used, represented by their proximity to the 45° line, proves that our AI-gram audibility model and the event gram are good predictors of the average normal listener score. The 120 Hz difference between optimal bandwidths for WN and SWN seems insignificant. The identification of an intermediate value for both noise spectra would be the basis of additional work. The difference in optimal AI thresholds T is likely due to the spectral emphasis of each noise (low frequencies in SWN and high frequencies in WN). The lower value obtained in WN could also be the result of other cues at lower frequencies, contributing to the score, when the burst is weak. More research will be needed to identify the optimal parameters and to precisely characterize the correlation between the scores and the event-gram model.

Figure 6(b) shows an event gram in SWN, for utterance f106ta, with the optimal bandwidth between the two horizontal dashed lines, leading to the identification of SNR_e. Below

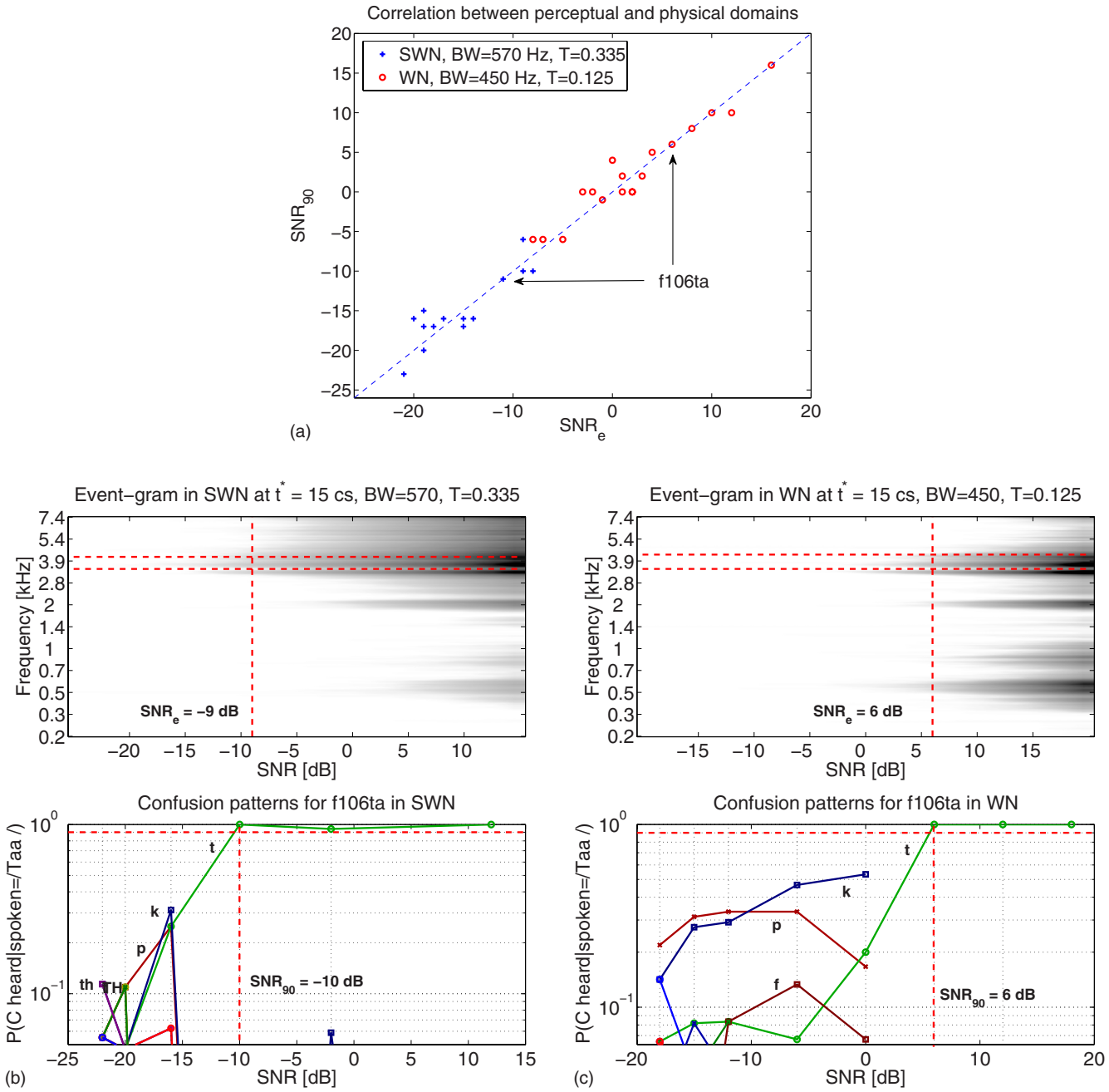

FIG. 6. (Color online) (a) Scatter plot of the event-gram thresholds SNR_e above 2 kHz, computed for the optimal burst bandwidth B having an AI density greater than the optimal threshold T compared to the SNR of 90% score. Utterances in SWN (+) are more robust than those in WN (○), accounting for the large spread in SNR. We can see that most utterances are close to the 45° line, showing the high correlation between the AI-gram audibility model (middle pane) and the event gram (right pane). The detection of the event-gram threshold SNR_e is shown on the event gram in SWN [top pane of (b)] and WN [top pane of (c)], between the two horizontal lines, for f106ta, and placed above their corresponding CPs. SNR_e is located at the lowest SNR where there is continuous energy above 2 kHz and spread in frequency with a width of B above AI threshold T. We can notice the effect of the noise spectrum on the event gram, accounting for the difference in robustness between WN and SWN.

are the CPs where $SNR_{90}=-10$ dB is shown by a vertical dashed line (thresholds are chosen in 1 dB steps, and the closest SNR integer value above 90% is chosen for SNR_{90}). Panel (c) shows the event gram and CPs for the same utterance in WN. The points corresponding to utterance f106ta are denoted by arrows.

Regardless of the noise type, we can see on the event grams the relation between the audibility of the 2–8 kHz range at t^* (in dark) and the correct recognition of /t/ even though the thresholds are much lower in SWN than in WN. More specifically, the strong masking of WN at high frequencies accounts for the early loss of the /t/ audibility as compared to SWN.

We conclude that the burst, as a high-frequency coinciding onset, is the main event accounting for the robustness of consonant /t/ independent of the noise spectrum. It presents different physical properties depending on the masker spectrum, but its audibility is strongly related to human responses in both cases. The event is therefore based on the audibility of the /t/ burst, not on any superthreshold property.

To further amplify the conclusions of the four-step method, we decided to run a psychophysical experiment where the /t/ burst would be truncated and study the resulting responses under lower noise conditions. We hypothesize that since the /t/ burst is the most robust-to-noise event, it is the strongest feature cuing the /t/ percept even at higher SNRs. A truncation experiment would therefore remove this crucial /t/ information.

IV. EXPERIMENT 2: TIME TRUNCATION

We next strengthen our conclusions drawn from Fig. 4 based on a confusion pattern and the event-gram analysis. Inspired by the work of Furui (1986), we truncated CV sounds in 5 ms steps and studied the resulting morphs. Our goal is to answer a fundamental research question raised by the four-step analysis of /t/: can the truncation of /t/ cause a morph to /p/, implying that the /t/ event is prefixed to consonant /p/, thereby showing that they share common features? Such a conclusion would be consistent with our observation that many /t/'s strongly morph to /p/ when the energy at high frequencies around t^* is masked by the noise.

A. Methods

Two SNR conditions, 0 and 12 dB SNRs, were used with SWN. The noise spectrum was identical to that used in PA07 (Phatak and Allen, 2007). In preliminary pilot studies, we identified the most common responses, and 22 were chosen that seem to accommodate our subjects. The final response task was then a forced choice among the 22 possible consonant CV responses or vowel only. A "vowel only" button was provided for those cases where no consonant could be identified. When queried, our subjects never expressed a need to add more response choices. Ten subjects participated in the final experiment.

1. Stimuli

The six tested CVs were /tɑ/, /pɑ/, /sɑ/, /zɑ/, /ʃɑ/, and /ʒɑ/ spoken by ten different talkers, for a total of 60 utterances. The times of the onset of each consonant and vowel were hand labeled. Onset truncations were generated automatically from these timing markers every 5 ms, including a "no truncation" condition and a "total" truncation condition, where only the vowel was played. One-half second of noise was prefixed to the CVs. The truncation was created by ramping the clean speech with half of a 10 ms Hamming window (5 ms ramp). The noise was added following the ramping. We only report the /t/ results here.

B. Results

The main conclusion of the /tɑ/ truncation experiment is the strong morph obtained for *all* of our stimuli when less than 30 ms of the /t/ burst are removed relative to the hand-labeled onset of the consonant. When presented with our truncated /tɑ/ sounds, listeners most commonly reported hearing /p/. Occasionally, /k/ or /h/ is reported, but when so, they have much lower average scores than /p/. In all cases, transitions took place within a 5–10 ms time frame. Specifically, the /t/ recognition abruptly dropped from above 90% at the onset of morphing to less than 10% within 5–10 ms.

Two main trends can be observed. Four out of ten utterances followed a hierarchical $/t/ \rightarrow /p/ \rightarrow /b/$ morphing pattern (group 1). In these cases, the consonant was first identified as /t/ for truncation times less than 30 ms; then /p/ is reported over a period spreading from 30 to 110 ms (one extreme case) and to finally reported as /b/.

The results (all four utterances) of group 1 are shown in Fig. 7. Note the significant variability in the crossover truncation times (the duration that the target and the morph scores overlap). This is due to both the natural variability in the /t/ burst duration and variation in the subject's responses. We have not analyzed the relative magnitudes of these two sources of variation. As a rule, the change in SNR from 12 to 0 dB had only a negligible impact on the scores (see Fig. 7).

The second trend (group 2) consisted of utterances that morph from $/t/ \rightarrow /p/$ but are also confused with /h/ and /k/. Five out of ten utterances (shown in Figs. 8 and 9) are in this group. The /h/ confusion is represented by the black dashed line and is stronger for the two top utterances, m102ta and m104ta [Figs. 8(a) and 8(b)]. A decrease in SNR from 12 to 0 dB caused a small increase in the /h/ score, almost bringing scores to chance performance (e.g., 50%) between those two consonants for the top two panels. This suggests a priming situation between these two sounds. The four lower panels show results for talkers m107 and m117, where a decrease in SNR causes a /k/ confusion which is as strong as the /h/ confusion but which differs from the 12 dB case, where competitor /k/ was not reported. The final example for the truncation of utterance f113ta (Fig. 9) shows a weak /h/ confusion to the /p/ morph, not significantly affected by a SNR change.

A noticeable difference of group 2 from group 1 is the absence of /b/ as a significant competitor. This discrepancy could simply be due to the limit of the truncation times. Utterances m104ta and m117ta [Figs. 8(b) and 8(d)] show weak /b/ confusions at the last truncation time tested. More test conditions would be needed to confirm this. Alternatively, since /p/ is a midstate between /t/ and /b/, we might simply study /pɑ/ truncations. In our /pɑ/ truncation data (not shown), not all /pɑ/s morph to /bɑ/; in most cases, no consonant is heard for the maximum truncation condition (i.e., only the vowel is reported). Therefore, if our $/t/ \rightarrow /p/$ results suggest that /p/ could be a precursor for /t/, we have not yet drawn this conclusion for $/p/ \rightarrow /b/$. More experiments will be required to resolve this apparent discrepancy. Of course, it is possible that the confusions with /h/ play a role in the loss of /b/ confusions.

In the figures, notice that for both groups 1 and 2, the onset time of the /t/ confusion depends on SNR. In most of the 0 dB case, the score for /t/ drops 5 ms earlier than for the 12 dB cases. This is likely due to masking of both sides of the burst energy, making them less audible and more difficult to be heard as an onset cue. When we study the AI grams for these cases, we see that the /t/-burst energy is weaker at t^*

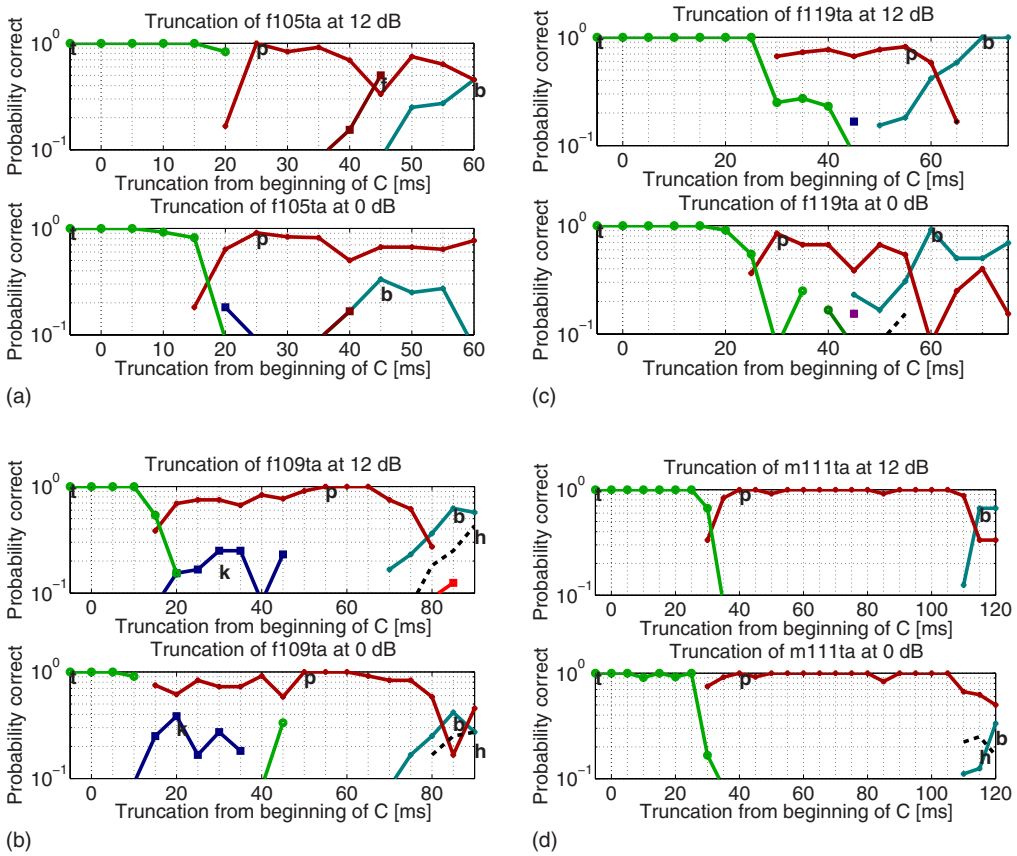

FIG. 7. (Color online) Group 1 utterances are defined as those which morph as /t/ → /p/ → /b/. For each panel, the top plot represents responses at 12 dB SNR and the lower those at 0 dB SNR. There is no significant SNR effect for these sounds. (a) Truncation of f105ta at 12 (top) and 0 dB SNRs (bottom). (b) Truncation of f109ta at 12 (top) and 0 dB SNRs (bottom). (c) Truncation of f119ta at 12 (top) and 0 dB SNRs (bottom). (d) Truncation of m111ta at 12 (top) and 0 dB SNRs (bottom).

(where the /t/ burst energy has its maximum). The best example of this somewhat subtle SNR effect is shown in Fig. 7(d).

As shown in Fig. 10, the pattern for the truncation of utterance m120ta was totally different from the other nine utterances included in the experiment. First, the score for /t/ did not decrease significantly after 30 ms of truncation. Second, /k/ confusions were present at 12 but not at 0 dB SNR, causing the /p/ score to reach 100% only at 0 dB. Third, the effect of SNR was pronounced. Figure 10 shows AI grams of this /tɑ/ at 12 (a) and 0 (b) dB SNRs. We can see that the burst is very strong for about 35 ms for both SNRs, which accounts for the high /t/ recognition in this range. For truncation times greater than 35 ms, /t/ is still identified with an average probability of 30%. This effect appears to be due to the level of high-frequency energy following the onset. As a result of the truncation, a coinciding onset of energy in the /t/-burst event frequency range is created, the duration of which is close to the natural /t/ burst.

For the 12 dB SNR case and for truncation times greater than 4 cs, this artificial /t/ burst appears to be weaker than the original strong onset burst. This explains the lower /t/ score compared to that in the untruncated version.

An unanticipated score inversion appears at 55 ms for the 0 dB SNR case. This /t/ peak is also weakly visible at 12 dB (left). We hypothesize that midfrequency energy, most likely around 0.7 kHz, is cuing /p/ at 12 dB but being masked at 0 dB SNR, enabling the /t/ recognition to rise. For the first 30 ms of truncation, this behavior is similar to that of the other utterances. The pattern observed for later truncation times is a demonstration of utterance variability, and can thus be explained without violating the /t/ burst event hypothesis. As shown by the truncation plots, when the /t/ burst is gone, beyond 80 ms, the /t/ score finally drops.

We conclude from the CV-truncation data that the *consonant duration* is an important timing cue used by listeners to distinguish /t/ from /p/. This duration depends on the natural duration of the /t/ burst. When only 5–10 ms more are truncated, the scores can drop dramatically. This demonstrates the high temporal sensitivity of the /t/ burst. As discussed earlier, we found in additional results (not shown) on truncated /pɑ/ that utterances are frequently confused with /bɑ/. This is consistent with the idea of a hierarchy of speech sounds, clearly present in our group 1 /tɑ/ examples. At the core of this hierarchy is /p/. By prefixing a short-duration burst, this can become /tɑ/ or /kɑ/. We have limited preliminary data that show that this works for several other vowels as well.

In summary, using our truncation procedure, we have independently verified that the high-frequency burst accounts for the noise-robust event corresponding to the discrimination between /t/ and /p/ even in moderate noisy conditions. Thus, we confirm that our approach of adding noise to iden-

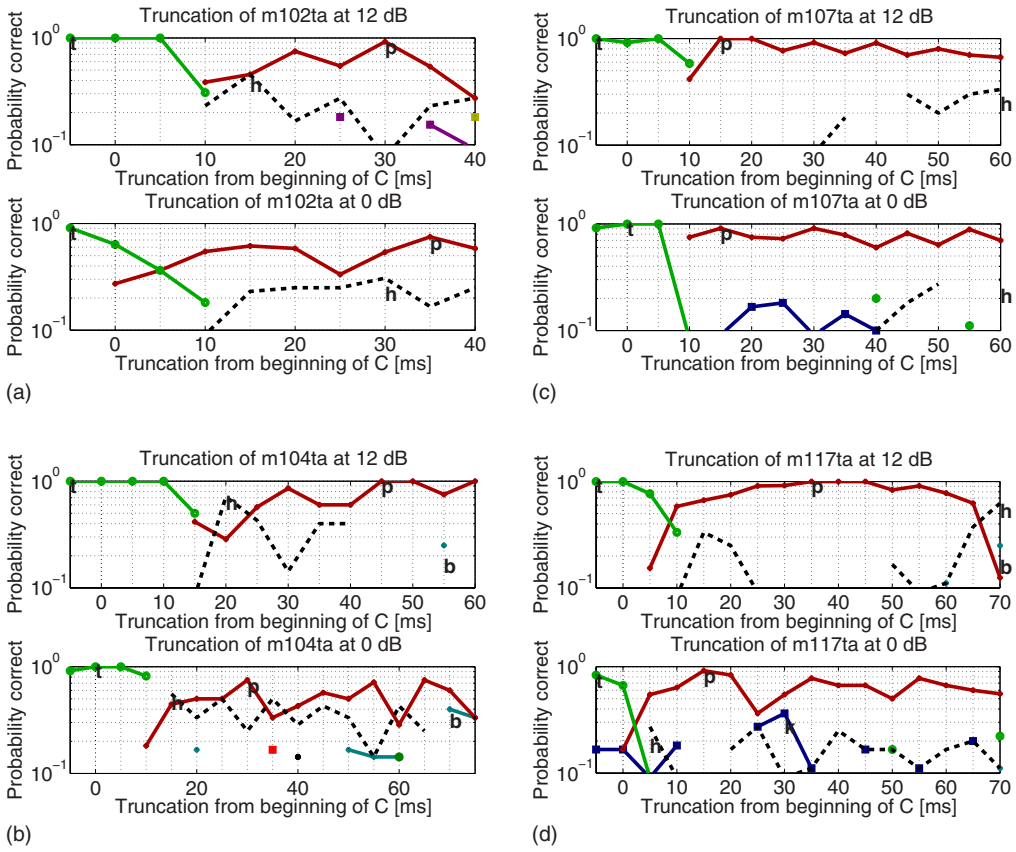

FIG. 8. (Color online) Utterances of group 2: Consonant /h/ strongly competes with /p/ (top), along with /k/ (bottom). For the top right and left panels: increasing the noise to 0 dB SNR causes an increase in the /h/ confusion in the /p/ morph range. For the two bottom utterances, decreasing the SNR causes a /k/ confusion that was nonexistent at 12 dB, equating the scores for competitors /k/ and /h/. (a) Truncation of m102ta at 12 (top) and 0 dB SNRs (bottom). (b) Truncation of m104ta at 12 (top) and 0 dB SNRs (bottom). (c) Truncation of m107ta at 12 (top) and 0 dB SNRs (bottom). (d) Truncation of m117ta at 12 (top) and 0 dB SNRs (bottom).

tify the most robust, and therefore crucial perceptual information, enables us to identify the primary feature responsible for the correct recognition of /t/.

C. Discussion

The results of our truncation experiment demonstrate that the /t/ recognition is masked after 30 ms for nine out of ten of our stimuli. This is in locked agreement with our analysis of the AI gram and event gram emphasized by our four-step analysis. We proved, therefore, that the leading edge of the /t/ burst from 2 to 8 kHz is used by our listeners to identify /t/ at all SNRs. We conclude that this burst is an across-frequency coincidence over a specific frequency range, and it plays the key role in the robust recognition of /t/.

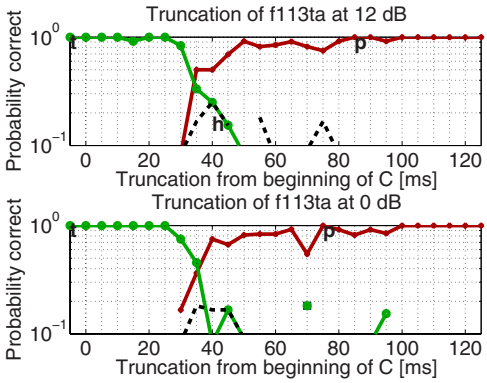

FIG. 9. (Color online) Truncation of f113ta at 12 (top) and 0 dB SNRs (bottom): Consonant /t/ morphs to /p/, which is slightly confused with /h/. There is no significant SNR effect.

Moreover, the /p/ morph that consistently occurs when the /t/ burst is truncated shows that consonants are not perceptual independent; thus, they share common cues. This hypothesis leads to the possible existence of "root" consonants. We presently view /p/ as a voiceless stop consonant root containing raw but crucial spectrotemporal information to which primary robust-to-noise cues can be added to form the various consonant of /p/'s confusion group. We have only demonstrated this here for the case of /t/. When CVs are mixed with masking noise morphing and priming are important empirical observations that are correlated with this conclusion.

Future work would need to verify that the /t/ recognition significantly drops when 30 ms of *only* the above 2 kHz burst region is removed. Such an experiment would further prove that it is an exclusively high-frequency /t/-burst event, making it not just sufficient but a necessary condition.

Dynamically modifying such timing events using signal processing techniques could lead to a new family of hearing

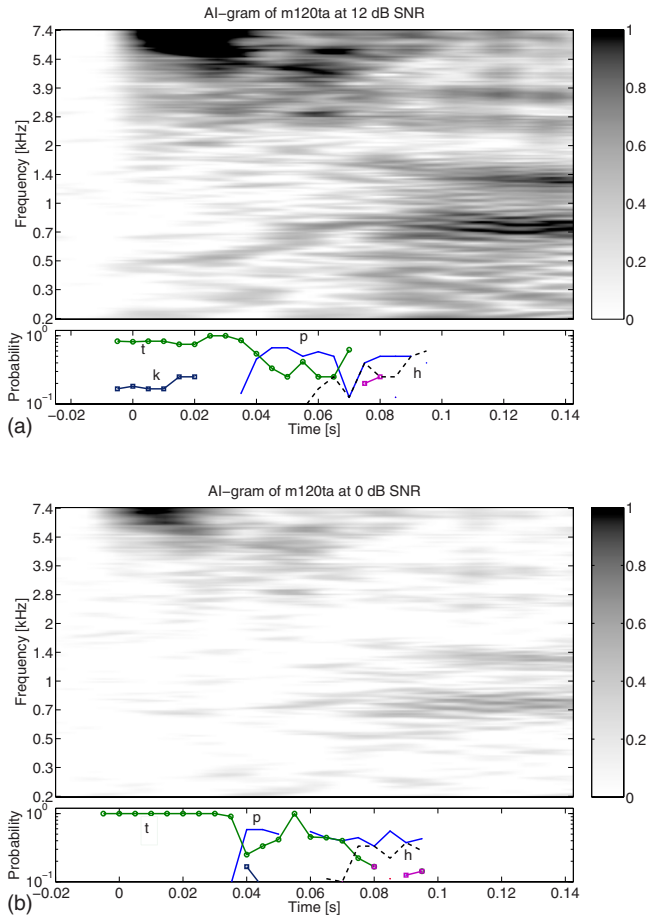

FIG. 10. (Color online) (a) 12 dB SNR zoomed AI gram. (b) 0 dB SNR zoomed AI gram. AI grams of m120ta, zoomed to a duration of 160 ms, in the consonant and transition regions at (a) 12 dB SNR and (b) 0 dB SNR. Below each AI gram are plotted the listener responses as a function of the truncation time, time synchronized. Uniquely for this utterance, the /t/ identification is still high after 30 ms of truncation, presumably because of the long-duration residual high-frequency (i.e., 2–8 kHz) energy. The target probability even overcomes the score for /p/ at 0 dB SNR at a truncation time of 55 ms, most likely because of a strong relative /p/ event present at 12 dB but weaker at 0 dB.

aids (Braida *et al.*, 1979), cochlear implants (Rabinowitz *et al.*, 1992; Shannon *et al.*, 1995), and robust automatic speech recognition (Hermansky, 1998).

Making automatic speech recognizers robust to noise, based on knowledge of human speech recognition, would be a tremendous breakthrough in an area where significant improvement is sorely needed (Lippmann, 1997; Dusan and Rabiner, 2005).

V. CONCLUSION

Our overall approach aims at directly relating our model of *speech audibility* in noise, the AI gram, a generalization of the AI, to the *confusion pattern discrimination measure* for consonant /t/. The AI gram represents an "input" measure, while the CPs are the "output" measure of the auditory speech processing system. Our approach is novel and represents a significant contribution toward solving the speech robustness problem as it has successfully led to the identification of the /t/ event as a synchronous 2–8 kHz temporal burst. This event is common across CVs identified with /t/ even if its physical properties widely vary across utterances. The strength of this acoustic feature leads to different degrees of robustness to noise. We have harnessed its natural variability to allow us to establish a correlation between the acoustic features and the scores. The correlation we observed between event-gram thresholds and 90% CP scores [Fig. 6(a)] confirms our hypothesis. This nearly perfect correlation between the /t/ burst strength and /t/ masked threshold scores has been demonstrated in a systematic manner across a large number of utterances.

Our results are consistent with the idea that the auditory brain is listening for onset transients. There is a small but significant literature on the importance of onset transients in both cortical research (Heil, 1997; Oertel, 2005; Heil, 2003; Shamma, 2003) and in the music perception research where timing onsets are known to carry important information about instrument identity.

Our onset-timing interpretation of speech cues resulted in a second experiment where we truncated the sounds from the onset and to the concept of a hierarchy of consonants forming the confusion group. It confirms our hypothesis that consonants forming a confusion group share common events. Finally, we have clearly shown that the /t/ burst is the primary feature for the identification of /t/ even in small amounts of noise. Primary events, along with a shared base of perceptual features, are used to discriminate consonants and characterize the consonant's degree of robustness to masking noise.

We have concluded that the event is therefore based on the audibility of the /t/ burst, not on any superthreshold property.

We believe that a hearing aid, tuned to such onset transients, could extract these cues and amplify them on a listener basis, resulting in significant improvement of speech identification in noisy environments for hearing impaired (HI) listeners.

ACKNOWLEDGMENTS

The authors wish to acknowledge the support of the Human Speech Recognition Group for the valuable comments and data collection. The authors would like to thank Etymotic Research and Starkey Laboratories for the financial support used to pay the subjects in the many experiments. Financial support was mostly provided by the ECE Department, UIUC. This work constitutes a portion of the first author's MS thesis.

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., and Neely, S. T. (**1997**). "Modeling the relation between the intensity JND and loudness for pure tones and wide-band noise," J. Acoust. Soc. Am. **102**, 3628–3646.
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.
Braida, L. D., Durlach, N. I., Lippmann, R. P., Hicks, B. L., Rabinowitz, W. M., and Reed, C. M. (**1979**). "Hearing aids: A review of past research on linear amplification, amplitude compression, and frequency lowering," American Speech and Hearing Association Monograph No. 19.
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**, 579–606.
Delattre, P., Liberman, A., and Cooper, F. (**1955**). "Acoustic loci and translational cues for consonants," J. Acoust. Soc. Am. **24**, 769–773 (Haskins work on painted speech).
Delgutte, B., Hammond, B., and Cariani, P. (**1998**). "Neural coding of the temporal envelope of speech: Relation to modulation transfer functions," in *Psychophysical and Physiological Advances in Hearing*, edited by A. Palmer, A. Reese, A. Summerfield, and R. Meddis (Whurr, London), pp. 596–603.
Dubno, J. R., Dirks, D., and Schaefer, A. (**1987**). "Effects of hearing loss on utilization of short-duration spectral cues in stop consonant recognition," J. Acoust. Soc. Am. **81**, 1940–1947.
Dubno, J. R., and Levitt, H. (**1981**). "Predicting consonant confusions from acoustic analysis," J. Acoust. Soc. Am. **69**, 249–261.
Dusan, S., and Rabiner, L. (**2005**). "Can automatic speech recognition learn more from human speech perception?," in *Trends in Speech Technology* Romanian Academy, pp. 21–36.
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).
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 — Analysis and preliminary ASR experiments," in Proceedings of the International Conference on Spoken Language Processing, (JEJU Island, Korea).
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.
Hant, J., and Alwan, A. (**2003**). "A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise," Speech Commun. **40**, 291–313.
Hawkins, S. (**2003**). "Roles and representations of systematic fine phonetic detail in speech understanding," J. Phonetics **31**, 373–405.
Heil, P. (**1997**). "Auditory cortical onset responses revisited. I. First-spike timing," J. Neurophysiol. **77**, 2616–2641.
Heil, P. (**2003**). "Coding of temporal onset envelope in the auditory system," Speech Commun. **41**, 123–134.
Hermansky, H. (**1998**). "Should recognizers have ears?," Speech Commun. **25**, 3–27.
Joris, P., Smith, P., and Yin, T. (**1998**). "Coincidence detection in the auditory system: 50 years after Jeffress," Neuron **21**, 1235–1238.
Kamm, C., Dirks, D., and Bell, T. (**1985**). "Speech recognition and the articulation index for normal and hearing-impaired listeners," J. Acoust. Soc. Am. **77**, 281–288.
Kewley-Port, D. (**1983**). "Time varying features as correlates of place of articulation in stop consonants," J. Acoust. Soc. Am. **73**, 322–335.
Klunder, K. R., Lotto, A. J., and Jenison, R. L. (**1995**). "Perception of voicing for syllable-initial stops at different intensities: Does synchrony capture signal voiceless stop consonants?," J. Acoust. Soc. Am. **97**, 2552–2567.
Lippmann, R. P. (**1997**). "Speech perception by humans and machines," Speech Commun. **22**, 1–15.
Lisker, L. (**1985**). "The pursuit of invariance in speech signals," J. Acoust. Soc. Am. **77**, 1199–1202.
Lobdell, B., and Allen, J. B. (**2006**). "An information theoretic tool for investigating speech perception," in Proceedings of Interspeech, p. 87.
Lobdell, B., and Allen, J. B. (**2007**). "Modeling and using the vu-meter (volume unit meter) with comparisons to root-mean-square speech levels," J. Acoust. Soc. Am. **121**, 279–285.
Lovitt, A., and Allen, J. (**2006**). "Using listener and talker variability to understand confusion patterns," in International Hearing Aid Research Conference.
Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.
Nguyen, N., and Hawkins, S. (**2003**). "Temporal integration in the perception of speech: Introduction," J. Phonetics **31**, 289–291 (overview of the special issue).
Oertel, D. (**2005**). "Importance of timing for understanding speech: Focus on Perceptual consequences of disrupted auditory nerve activity," J. Neurophysiol. **93**, 3044–3045 (editorial focus).
Phatak, S., and Allen, J. B. (**2007**). "Consonant and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am. **121**, 2312–2326.
Rabinowitz, W., Eddington, D., Delhome, L., and Cuneo, P. (**1992**). "Relations among different measures of speech reception in subjects using a cochlear implant," J. Acoust. Soc. Am. **92**, 1869–1881.
Régnier, M., and Allen, J. B. (**2007a**). "Perceptual cues of some CV sounds studied in noise," *Abstracts* (AAS, Scottsdale).
Régnier, M., and Allen, J. B. (**2007b**). "The importance of across frequency timing coincidences in the perception of some English consonants in noise," *Abstracts* (ARO, Denver, CO).
Repp, B., Liberman, A., Eccardt, T., and Pesetsky, D. (**1978**). "Perceptual integration of acoustic cues for stop, fricative, and affricate manner," J. Exp. Psychol. **4**, 621–637.
Rhebergen, K., Versfeld, N., and Dreschler, A. (**2006**). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," J. Acoust. Soc. Am. **120**, 3988–3997.
Shamma, S. (**2003**). "Physiological foundations of temporal integration in the perception of speech," J. Phonetics **31**, 495–501.
Shannon, C. E. (**1948**). "The mathematical theory of communication," Bell Syst. Tech. J. **27**, 379–423; 1948) 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., and Klatt, D. (**1974**). "Role of formant transitions in the voiced-voiceless distinstion for stops," J. Acoust. Soc. Am. **55**, 653–659.
Strope, B., and Alwan, A. (**1997**). "A model of dynamic auditory perception and its application to robust word recognition," IEEE Trans. Acoust., Speech, Signal Process. **5**, 451–464.
Summerfield, Q., and Haggard, M. (**1977**). "On the dissociation of spectral and temporal cues to the voicing distinction in the initial stop consonants," J. Acoust. Soc. Am. **62**, 51–61.