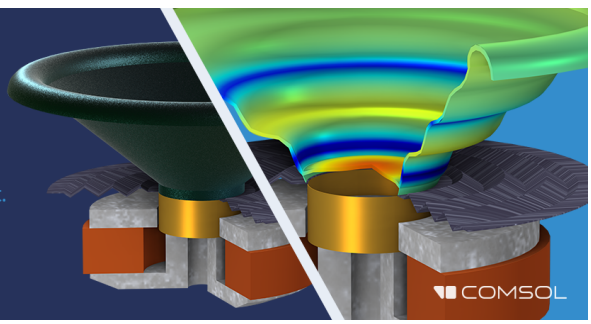
AUGUST 26 2019

A framework for labeling speech with acoustic cues to linguistic distinctive features

Shreya Huilgol; Jinwoo Baik; Stefanie Shattuck-Hufnagel

J Acoust Soc Am 146, EL184–EL190 (2019)
https://doi.org/10.1121/1.5121717
CHORUS



A framework for labeling speech with acoustic cues to linguistic distinctive features

Shreya Huilgol,[a] Jinwoo Baik, and Stefanie Shattuck-Hufnagel
Speech Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology, 50 Vassar Street, Cambridge, Massachusetts 02139, USA
shuilgol@wellesley.edu, jinwoobaik1@gmail.com, sshuf@mit.edu

Abstract: Acoustic cues are characteristic patterns in the speech signal that provide lexical, prosodic, or additional information, such as speaker identity. In particular, acoustic cues related to linguistic distinctive features can be extracted and marked from the speech signal. These acoustic cues can be used to infer the intended underlying phoneme sequence in an utterance. This study describes a framework for labeling acoustic cues in speech, including a suite of canonical cue prediction algorithms that facilitates manual labeling and provides a standard for analyzing variations in the surface realizations. A brief examination of subsets of annotated speech data shows that labeling acoustic cues opens the possibility of detailed analyses of cue modification patterns in speech.

[DDO]
Date Received: April 14, 2019 **Date Accepted:** July 24, 2019

1. Introduction

The MIT Speech Communication Group has developed a framework for labeling speech with acoustic cues related to distinctive features. *Acoustic cues* are physically observable patterns in the speech signal that can be extracted and interpreted to provide information about the speaker, or about the underlying message. Most often, the underlying message includes a sequence of *words*. The collection of words in a language is the *lexicon* (or dictionary), and each word within it is associated with a sequence of *phonemes*. Phonemes comprise the inventory of contrastive speech sounds in that language, and are denoted within slashes, e.g., /ih/. Each phoneme can be described in further detail in terms of *distinctive features*, such as whether it is a vowel or a consonant, or whether it should be produced with the lips. Distinctive features and their values are denoted with brackets, e.g., [+vowel]. As an example, the phoneme/ih/can be described as [+vowel, +high, −back]; a description of the phoneme /b/ is [+consonantal, −sonorant, −continuant, +labial, +voiced].

In the framework described in this paper, it is proposed that the abstract, symbolic hierarchy of distinctive features, phonemes, and words, is not directly extractable from the speech signal. Instead, only acoustic cues can be observed. These acoustic cues can then be used to infer the underlying distinctive features, and in turn, the associated phonemes, and the words. This approach aims to make clear the distinction between the physically observable acoustic cues versus the underlying abstract linguistic units. This departs from current database annotation schemes, which mark speech corpora with word and phone labels (e.g., the TIMIT database, Garofolo *et al.*,1993). *Phones* are defined as intervals in the speech signal that correspond to individual speech sounds, and are denoted within brackets, e.g., [ih]. (Instead of IPA symbols, the ARPABET notation is used here to facilitate discussion of phone labels in the TIMIT database.) In this typical labeling framework, the physical speech interval is considered to correspond to the abstract category; the constituent parts of the speech signal are not examined separately. In this paper, we describe a transcription approach based on the explicit labeling of individual acoustic cues to the underlying distinctive features. The motivation behind the development of this framework is described in Sec. 2, the specific procedure that is used for manually labeling acoustic cues in speech is outlined in Sec. 3, and a discussion of how the resulting labeled speech allows for more insightful speech analysis is given in Sec. 4.

2. Background

There are three important and well-known issues that arise in the phone-based labeling framework. The first is the wide variation in acoustic characteristics for speech intervals that are

[a]Author to whom correspondence should be addressed.

assigned the same phone label. For example, speech intervals that are labeled [ih] may have very different formant patterns, depending on the characteristics of adjacent regions. In most applications, these variations are considered to be noise components (despite their predictability from context), and are not analyzed in detail or used to aid in speech recognition or analysis.

In contrast, in the proposed framework such systematic contextual variability is considered useful information: the various surface realizations of the same underlying phoneme sequence are seen as productions of different combinations of cues drawn from the set of relevant acoustic cues for each feature. The various cue modification patterns can be catalogued, and this becomes a source for a richer description of the produced speech (similar to Schuppler *et al.*, 2011). At the same time, this approach usefully reduces the rich information that the signal provides by focusing on the acoustic characteristics which are directly related to the linguistic feature contrasts, only extracting the information that is most relevant for finding those distinctive features. While both the feature cue patterns and other information in the signal can be useful for recognizing other aspects of what is signaled by a spoken utterance (such as the individual speaker, the speaker's tribal affiliation, the speaker's physiological, or attitudinal state, etc.), the explicit extraction of cues to the distinctive features of the intended utterance provides the information that the listener needs in order to access the intended words in the lexicon.

The second issue concerns the timing of acoustic cues relative to the starts and ends of phone intervals. An interval assigned a phone label often does not include all the information for that speech sound, which may instead spill over to intervals assigned to adjacent speech sounds; on the other hand, a phone interval often includes information about adjacent or neighboring sounds. For example, the formant pattern at the beginning of an [ih] region in an utterance of the word *bit* can be heavily influenced by the preceding consonant, providing information about that consonant. As a result, listeners may hear a /b/ onset consonant when playing back the sound interval of phonation for the vowel phone because the formant transitions at the onset of that vocalic region signal a labial consonant. This information is lost in a phone-based transcription but captured in a feature-cue-based transcription. Also, individual feature cues can be dropped or added separately; for example, obstruent consonants may sometimes show a clear closure without a release or vice versa. Without individual labels for these events, it is not possible to determine whether such patterns are systematic with respect to context, which may be useful information (Schuppler *et al.*, 2011). Furthermore, the phone-based method does not capture tokens which are produced with a combination of attributes that are characteristic of different phones. An example would be the case of a flapped /t/, which in some cases retains a strong closure and/or release of very short duration, and in other cases becomes almost glide-like, with just a small dip in the amplitude of the waveform (Yun *et al.*, 2017). For such cases, a phone-based annotation scheme forces a choice among the available phone-category labels, e.g., stop [t] or a flap [dx], while a feature-cue-based annotation captures the relevant acoustic characteristics.

The proposed feature-cue-based transcription takes advantage of the fact that the acoustic patterns associated with a speech sound can be separated into identifiable characteristics which are not bound to phone intervals. For example, as noted above, the formant transition patterns at the start of phonation for a vowel can provide place of articulation information for a previous consonant, and at the same time, the bandwidth of the first formant, along with the incidence of extra poles and zeros in the first formant region, can provide information about whether nasalization is present. Such patterns can occur within the time interval traditionally associated with a vowel, but can provide information about sounds that are outside of that time interval, namely, a previous obstruent consonant, or a following nasal consonant. Because a cue-based annotation system can indicate where an individual feature cue begins and ends in the signal, it can capture such patterns. In sum, patterns such as overlapping acoustic cues in the same region, acoustic cues to the same distinctive feature that spread over adjacent regions, and combinations of individual acoustic cues, even those that do not correspond to a traditional phone, can be described easily in the proposed feature-cue-based system.

The third issue is how to define the speech sounds corresponding to a word. For example, in TIMIT, the word "water" is often labeled as [w ao dx axr]. However, it is also possible to label it with the sequence [w ao tcl t er]. Because the set of potential phones is not related to contrasts in meaning (as with phonemes), but can vary with the level of detail desired, there is no standard way to assign a sequence of speech sounds to a word, with phone labels. This becomes even more problematic if modifications and/or cross-word boundary effects are involved. For example, a speech interval corresponding to the words "had your" was labeled as [hv eh dcl jh ih] in the TIMIT database. In this instance, the phone sequence for the word "your" becomes [jh ih], which is not considered a standard pronunciation for the word.

The approach proposed in this paper resolves this issue by separating the varying surface pronunciations of a word from its underlying sequence of non-varying phonemes. A phoneme sequence that defines a given word (and distinguishes it from other words in the language) is fixed, while the surface realization in the form of phones is highly variable. This separation is useful because the relationship between the sequence of phones in the surface production and the underlying phonemic sequence is not always apparent, even if sub-segmental cues are used (e.g., Schuppler *et al.*, 2009). The central goal of speech recognition for both human and machine listeners is to identify the underlying sequence of contrastive phonemes, which in turn enables identification of the word sequence, but in many approaches the identification of the underlying phoneme sequence has been conflated with identification of the surface phone sequence. For many cases, the distinction between these two goals is minor: it is readily apparent that the phoneme sequence /ih ch/ for the word "itch" is highly likely to be realized as the phone sequence [ih tcl ch]. However, in other cases the distinction is major. For example, consider the word sequence *did you eat yet*, which can be produced in a severely reduced fashion, in a form which is suggested by the phone sequence [jh ih tcl ch eh], This phone sequence would not be immediately recognizable as a surface rendition of the underlying phoneme sequence of /d ih d y uw iy t y eh t/. Such examples of extreme reduction (Johnson, 2004; Schuppler *et al.*, 2011) abound, as when the word sequence "do you have" can be produced as something like "jev."

The challenging issues that arise from phone-based labeling have led to previous work on transcribing with sub-phonetic units. Livescu *et al.* (2007) describes a framework for labeling articulatory gestures which also uses multi-tier labels and takes advantage of the tight link between changing articulatory configurations and their acoustic consequences. However, the distinction between the surface and underlying descriptions was not made, and the labels themselves are in the form of discrete time intervals so that the issues discussed above are still problematic: it remains challenging to capture patterns of overlapping feature cues and their deletion, addition, and recombination. The examples cited earlier show that it is useful to make clear the distinction among (1) the underlying phonemes, which are nominally fixed for a given word, (2) their surface realizations, which are observable as combinations of acoustic cues in the speech signal, and (3) the common combinations of such acoustic cues, which are called phones. The proposed framework shows that capturing the relationship between the "ground truth" phonological representation of words and fine-grained sub-segmental descriptions is possible, by using the phonemes-to-distinctive-features relation within the abstract domain, and the link between the distinctive feature values and the related acoustic cues across the abstract and physical domains.

Adopting a feature-cue-based system for transcribing and analysing the speech signal has application to automatic speech recognition (ASR), and also a number of additional advantages. By re-framing the question of how listeners represent the initial analysis of a spoken utterance as the question of how listeners extract individual cues to distinctive features, this approach enables the exploration of testable models of human speech perception, an important endeavor which has become less tenable over recent years as ASR models have moved away from the goal of modeling human speech processing. To the extent that individual-feature-cue-based processing provides a model of human speech processing, such an approach has the potential to shed light on the processes of speech and language development in children, and of second language learning, as well as providing the basis for improved understanding of the mechanisms underlying clinical speech disabilities; this could support the development of improved intervention methods. Thus, testing the feature-cue-based approach is arguably a critical next step in evaluating theories and models of speech processing.

3. Labeling speech with acoustic cues to distinctive features

In this paper, we focus on the acoustic cues that are related to the speaker's intended distinctive features of manner, place, and voicing, and can be used to infer the phonemes and the associated word sequence. The manner features can be termed articulator-free features, while the place, nasalization, and voicing features can be called the articulator-bound features (Keyser and Stevens, 1994). These classes correspond to Halle's (1992) distinction between articulator-free features, like [consonantal], which do not specify an articulator, and articulator-bound features, like [labial], which do. Acoustic cues to features can be interpreted as evidence for a sequence of feature bundles (i.e., sets of distinctive features that can be interpreted together as evidence for a phoneme) and thus act as the earliest abstract units related to segmental units (Stevens, 2002). A list of acoustic cues to distinctive features and their related feature values is shown in the table in the supplementary material.[1]

Corresponding to the two groups of distinctive features, there are two broad classes of acoustic cues to the distinctive features: *Landmarks* (LMs) and *Other Acoustic Cues* (OACs). Acoustic LMs are points in the speech signal where acoustic cues are most evident, and information about distinctive features are easily extracted (Stevens, 2002). A LM can be described as an abrupt spectral change that signals one of the manner features of an underlying phoneme; such discrete events include maxima of vowels, minima of glides, and closures and releases of stop, fricative, and nasal consonants. The remaining class, OACs, is related to place, nasalization, and voicing features. These cues extend over intervals of time, can be found in the vicinity of the LMs, and further specify distinctive features. By combining the time-point annotation of LMs with the time-interval annotation of OACs, this system captures cue patterns that spread beyond the inter-LM region associated with a particular phoneme. It can handle cases where the LMs for a phoneme are missing, e.g., a voiceless coda consonant without closure or release LMs, cued primarily by the duration of the preceding vowel, or even more extreme cases of massive reduction (Johnson, 2004; Niebuhr and Kohler, 2011) where phone-based annotation is impractical.

In the proposed feature-cue-based labeling system for English, there are eight types of LM acoustic cues: ⟨V⟩ (vowel), ⟨G⟩ (glide), ⟨Nc⟩ (nasal consonant closure), ⟨Nr⟩ (nasal consonant release), ⟨Fc⟩ (fricative consonant closure), ⟨Fr⟩ (fricative consonant release), (stop consonant closure), and ⟨Sr⟩ (stop consonant release). Acoustic cues are indicated by chevrons, e.g., ⟨V⟩. Some LM acoustic cues can provide evidence for values for sets of distinctive features: a nasal consonant closure ⟨Nc⟩ would indicate the following features and values: [+consonantal, +sonorant, −continuant]. The same set would be indicated for a ⟨Nr⟩ LM acoustic cue. Similarly, ⟨Fc⟩ or ⟨Fr⟩ provides evidence for [+consonantal, −sonorant, +continuant], and or ⟨Sr⟩ for [+consonantal, −sonorant, −continuant].

The list of OACs is longer: there are 32 OACs in the current scheme. Several pertain to vowel and glide identity: ⟨high⟩, ⟨mid⟩, and ⟨low⟩ (acoustic cues indicated by *F*1 location in the frequency domain, for the distinctive feature sets [+high], [−high, −low], and [+low], respectively); ⟨front⟩ and ⟨back⟩ (acoustic cues describing *F*2 location, for the features [−back], and [+back], respectively); ⟨atr⟩ and ⟨ctr⟩ (acoustic cues for relative *F*1 and *F*2 extremities, for the features [+advanced tongue root], and [+constricted tongue root], respectively; and ⟨lat⟩ and ⟨rhot⟩ (acoustic cues for *F*3 location, for the features [+lateral], and [+rhotic], respectively). (In other descriptions of distinctive features, an alternative notation is possible that replaces [+atr] and [+ctr] with [+tense], without distinguishing between them.)

The acoustic cues to consonant place have a two-part notation, e.g., ⟨lab-FTc⟩. The first part denotes the place of articulation for which the acoustic cue provides evidence. In the present scheme, five labels are used: ⟨lab⟩, ⟨den⟩, ⟨alv⟩, ⟨pal⟩, and ⟨vel⟩, for acoustic cues that indicate the distinctive feature values [+labial], [+dental], [+alveolar], [+palatal], and [+velar]. (In an equivalent alternative description, the corresponding distinctive feature sets could be [+lips], [+tongue blade, +anterior, +distributed], [+tongue blade, +anterior, -distributed], [+tongue blade, −anterior, +distributed], and [+tongue body].) The second part of the notation points to the location in the signal for the acoustic cue. Formant transitions occur at the edges of vowels and consonant closure intervals; closures occur during the vowel-to-consonant transition and releases occur during the consonant-to vowel-transition (Story and Bunton, 2010). Thus, ⟨FTc⟩ denotes an acoustic cue found during the formant transition leading to the consonant closure, and ⟨FTr⟩ denotes a cue at the formant transition following a consonant release. The ⟨SB⟩ notation denotes acoustic cues found at frication or the release burst of a consonant.

Onset of nasalization, detected by increased bandwidth in the first formant region, with possible extra pole-zero pairs in the low frequencies, is marked ⟨n, and the offset is marked n⟩; these acoustic cues are related to the features [+nasal] and [−nasal]. The beginning of glottal voicing is labeled with ⟨g and the end with g⟩. These acoustic cues can be interpreted, in conjunction with OACs, to infer the distinctive features [+voiced] or [−voiced]. The onset and offset of aspiration (noise distributed throughout the frequency spectrum, with or without observable formant structure) are marked ⟨h, and h⟩, respectively, and this set of cues is related to the feature [spread glottis]. Irregular pitch periods, also described as glottalization, are marked with ⟨ipp at onset and ipp⟩ at offset. The acoustic cue of irregular pitch periods is interpreted contextually: it provides evidence for the distinctive feature [+constricted glottis] for glottalized /t/, or can signal prosodic events, such as intonational phrase boundaries or a phrasally-prominent syllable (Pierrehumbert and Talkin, 1992; Dilley *et al.*, 1996; Garellek, 2014).

3.1 Procedure for manual labeling of speech files

The procedure for labeling speech files with acoustic cues is shown in Fig. 1. It is assumed that the utterance has the word transcription provided. A Praat TextGrid file is created, with the word sequence entered in the first tier called *words*. A series of programs (coded in Java) are then used to prepare the TextGrid file for labeling. When the hand labels are completed, additional programs are used to check the labels and further analysis is carried out. The provided word sequence is first checked with a comprehensive dictionary, such as the Carnegie Mellon University online dictionary and/or the TIMIT dictionary, to ensure that all words in the TextGrid file are valid and have corresponding phone sequences listed (Check Words program). Next, assuming all words are found in the dictionaries, the associated dictionary-generated phone sequences are converted into standard phoneme sequences, which are placed in a new TextGrid tier called *phonemes* (Generate Phonemes). Once the phonemes have been generated, canonical cue prediction programs create new tiers titled *predLM*, *pred_vgplace*, *pred_cplace*, *pred_nasal*, and *pred_glottal*, with their respective predicted canonical acoustic cues in the form of LM cues, vowel and glide place cues, consonant place cues, nasalization cues, and glottal cues (Generate predLM Cues, Generate pred_vgplace Cues, …, Generate pred_glottal Cues); it is also possible to generate all predicted cue tiers at once (Generate All Prediction Cues). Additional tiers of particular interest in analysis are added, e.g., LMmods (for separate markings of LM deletions and insertions, and Comments.

The process described thus far prepares a sound file for manual annotation. The labeler uses the automatically generated information to facilitate annotation moving labels to their realized locations (for cues that are realized as predicted), deleting predicted labels that are not realized in the signal, and adding labels for cues not predicted from the underlying phonemes.

Once manual labeling is complete two utility programs are used to check TextGrid files for errors and/or to convert them for further use. The first is a Diagnostic Check program that checks for errors in the tiers and labels of TextGrid files, flagging any labels not in the acceptable list, for manual correction. Second, the Text Conversion program converts a Praat TextGrid file into a plain-text file. For each label in a tier, the time information of the label, the label name, and the tier name are output on separate lines, and the final result is printed to a text file that is readable by spreadsheet programs. We turn now to the labeling tiers themselves.

3.2 Seven tiers for labeling acoustic cues

The LM tier contains the LM labels of the speech file. As described previously, the list of possible labels is: ⟨V⟩, ⟨G⟩, ⟨Nc⟩, ⟨Nr⟩, ⟨Fc⟩, ⟨Fr⟩,⟨Sc⟩, and ⟨Sr⟩. All LM cues that appear in the speech signal that were also predicted from the word transcription, via the standard phoneme sequence, are positioned in the LM tier. The LMmods tier contains LM modifications. Deviations from predicted LMs, i.e., deletions and insertions, are marked in the LMmods tier, with "−x" for deletion and "−+" for insertion. For example, if a predicted V LM is not observed in the speech signal, it is removed from the LM tier, and a "⟨V⟩-x" label is marked in the LMmods tier.

The vgplace tier includes the place of articulation information for all vowel and glide LMs. The vgplace labels include: ⟨high⟩, ⟨mid⟩, ⟨low⟩, ⟨front⟩, ⟨back⟩, ⟨atr⟩, ⟨ctr⟩, ⟨lat⟩, and ⟨rhot⟩. The cplace tier contains information about the place of articulation for all consonant formant transitions and burst/frication spectra, with labels such as ⟨lab-FTc⟩, ⟨alv-SB⟩, etc. The nasal tier marks the intervals of nasalization with ⟨n and n⟩ labels. The glottal tier marks the acoustic cues which indicate the configuration of the glottal region, with labels such as ⟨g, h⟩, ⟨ipp, etc. The vgplace, cplace, nasal, and glottal tiers provide labels for the OACs. These tiers do not have an associated modifications tier: the predicted and realized cues are positioned at their locations, unrealized cues are simply removed, and inserted cues are added, all in the same tier. Finally, the comments tier includes notes on labels in the LMmods or vgplace tiers. An example of speech labeled with acoustic cues, including LM modifications, is shown in Fig. 2.

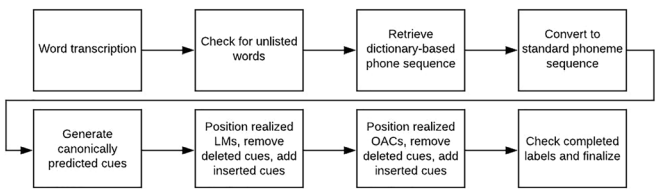


Fig. 1. Eight step process to producing speech files labeled with acoustic cues.

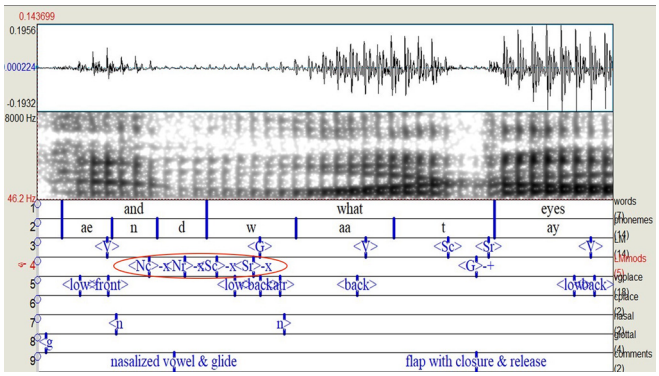


Fig. 2. (Color online) Praat Textgrid that displays the acoustic cue labels for the word sequence "… and what eyes …" from the TIMIT database. Note the labels in the fourth tier (LMmods tier) that show modifications from the predicted LM acoustic cues. The closures and releases for the/n/and/d/in "and" are deleted. Also, the/t/ in "what" has an added glide LM ⟨G⟩−+ in the fourth tier (LMmods tier), along with the predicted stop closure LM and stop release LM ⟨Sr⟩ in the third tier (LM tier).

3.3 Speech databases labeled with acoustic cues

Using the acoustic cue labeling process described above, three databases have been labeled with acoustic cues. The LAFF VCV database (http://hdl.handle.net/1721.1/25142) contains vowel-consonant-vowel utterances spoken by 3 speakers (1 F, 2 M), for 6 vowels and 26 glides/consonants. The UConn Isolated Words database contains speech from 4 speakers, each producing about 1200 commonly used monosyllabic words. Speech for 2 speakers (1 F, 1 M) has been labeled. Finally, a subset of the TIMIT database was also labeled for 10 sentences each produced by 2 speakers (1 F, 1 M) from 8 dialect regions in the TRAIN Development set, for a total of 160 sentences. Table 1 shows the occurrences of acoustic cue labels from subsets of each of these databases. For canonical utterances, the method requires about 500× real time; for more modified speech, the method can take up to 1200× real time, similar to the approach described in Livescu *et al.* (2007).

Note that there are more vowel LMs ⟨V⟩, stop consonant closure ⟨Sc⟩, and stop release ⟨Sr⟩ LMs than other types of LMs. Also, there are fewer modifications (i.e., deletions, insertions) in the LAFF VCV and UConn Isolated Words databases, compared to the continuous speech TIMIT sentences. In the Isolated Words and TIMIT, most of the fricative consonant closure ⟨Fc⟩ and fricative consonant release LMs ⟨Fr⟩ are realized, in contrast to the stop consonant closure and stop consonant release ⟨Sr⟩ LMs, which have more deletions. For the three databases, nasal consonant

Table 1. Counts of LM acoustic cue labels for subsets of LAFF VCV, UConn Isolated Words, and TIMIT databases. Predicted, realized, deleted, and inserted acoustic cues are listed in rows marked pred, real, del, and ins, respectively. Numbers in parentheses are % in terms of predicted cues. Counts are given for Vowel ⟨V⟩, Glide ⟨G⟩, Nasal closure ⟨Nc⟩, Nasal release ⟨Nr⟩, Fricative closure ⟨Fc⟩, Fricative release ⟨Fr⟩, Stop closure ⟨Sc⟩ , Stop release ⟨Sr⟩, and all LM acoustic cues.

		⟨V⟩	⟨G⟩	⟨Nc⟩	⟨Nr⟩	⟨Fc⟩	⟨Fr⟩		⟨Sr⟩	all LMs
LAFF VCV	pred	1984	108	81	81	340	340	416	416	3676
	real	1887	107	81	81	340	340	407	407	3650
		(99.6)	(99.1)	(100)	(100)	(100)	(100)	(97.8)	(97.8)	(99.3)
	del	7 (0.4)	1 (0.9)	0 (0)	0 (0)	0 (0)	0 (0)	9 (2.2)	9 (2.2)	26 (0.7)
	ins	0 (0)	9 (2.6)	0 (0)	0 (0)	4 (1.2)	8 (2.4)	4 (1.0)	4 (1.0)	29 (0.8)
UConn Isol	pred	463	343	155	155	262	262	474	474	2588
Words	real	459	338	152	148	259	261	363	441	2421
		(99.1)	(98.5)	(98.1)	(95.5)	(98.9)	(99.6)	(76.6)	(93.3)	(93.5)
	del	4 (0.9)	5 (1.5)	3 (1.9)	7 (4.5)	3 (1.1)	1 (0.4)	111 (23.4)	33 (7.0)	167 (6.5)
	ins	6 (1.3)	11 (3.2)	0 (0)	0 (0)	3 (1.1)	4 (1.5)	1 (0.2)	1 (0.2)	26 (1.0)
TIMIT	pred	2038	971	517	517	804	804	1091	1091	7833
TRAIN	real	1986	899	467	461	722	741	804	759	6839
Dev Set		(97.4)	(92.6)	(90.3)	(89.2)	(89.8)	(92.2)	(73.7)	(69.6)	(87.3)
	del	52	72	50	56	82	63	287	332	994
		(2.6))	(7.4)	(9.7)	(10.8)	(10.2)	(7.8)	(26.3)	(30.4)	(12.7)
	ins	9	156	6	7	25	23	37	54	317
		(0.4)	(16.1)	(1.2)	(1.4)	(3.1)	(2.9)	(3.4)	(4.9)	(4.0)

closure ⟨Nc⟩ and nasal consonant release ⟨Nr⟩ LMs are unlikely to be inserted. These preliminary observations show that the acoustic cue labeling framework enables detailed observations of cue modifications in speech.

4. Discussion

The acoustic cue labeling system and cue-labeled databases described in this paper provide information that is not immediately obvious in phone-labeled speech. For example, it can capture whether a stop was produced with both an abrupt closure and release, and whether a glottalized /t/ also shows evidence of an oral closure and release. In this sense, it is a robust way of labeling the information in an utterance that reflects the speaker's intended speech sounds while requiring few arbitrary decisions by the labeler. By capturing the LMs and OACs to features, this labeling system facilitates study of the range of ways that a phoneme can be produced, including information that may not be available from a phone-based annotation. For example, it specifies, in the LM modification tier, information about where an expected LM type is replaced by a different type. An example would be a glide-like flapped /t/, instead of a /t/ closure followed by a /t/ release. In this case, the LMmods labels −x, ⟨G⟩−+, ⟨Sr⟩−x, reflect the deletion of the abrupt closure and release LM cues, and the production of a glide-like cue instead. This method of labeling allows for less ambiguity in descriptions and a more precise and accurate way of characterizing the sound produced, compared to currently widely-used labeling schemes using phones. Finally, preliminary tabulations show that labeling the acoustic cues enables more detailed analysis of the cue modifications that occur in speech.

Acknowledgments

This research was supported in part by NSF Grant Nos. IIS 1651190 and BCS 1827598, and by the UROP program at MIT. The authors thank Jeung-Yoon Choi for guidance and support in developing the labeling system and in the preparation of the manuscript, and Christine Soh for help with database analysis.

References and links

[1]See supplementary material at https://doi.org/10.1121/1.5121717 for a list of acoustic cues to distinctive features and their related feature values.

Dilley, L. C., Shattuck-Hufnagel, S., and Ostendorf, M. (**1996**). "Glottalization of vowel-initial syllables as a function of prosodic structure," J. Phonetics **24**, 423–444.

Garellek, M. (**2014**). "Voice quality strengthening and glottalization," J. Phonetics **45**(1), 106–113.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (**1993**). "The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM," Linguistic Data Consortium, pp. 207–212.

Halle, M. (**1992**). "Features," in *The Oxford International Encyclopedia of Linguistics*, edited by W. Bright (Oxford University Press, New York).

Johnson, K. (**2004**). "Massive reduction in conversational American English," in *Spontaneous Speech: Data and Analysis*, in *Proceedings of the 1st Session of the 10th International Symposium*, pp. 29–54.

Keyser, S. J., and Stevens, K. N. (**1994**). "Feature geometry and the vocal tract," Phonology **11**, 207–236.

Livescu, K., Bezman, A., Borges, N., Yung, L., Cetin, O., Frankel, J., and Lavoie, L. (**2007**). "Manual transcription of conversational speech at the articulatory feature level," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, Vol. 4, pp. IV-953.

Niebuhr, O., and Kohler, K. (**2011**). "Perception of phonetic detail in the identification of highly reduced words," J. Phonetics **39**, 319–329.

Pierrehumbert, J., and Talkin, D. (**1992**). "Lenition of /h/ and glottal stop," in *Papers in Laboratory Phonology II* (Cambridge University Press, Cambridge, UK), pp. 90–117.

Schuppler, B., Dommelen, W. V., Koreman, J., and Ernestus, M. (**2009**). "Word-final [t]-deletion: An analysis on the segmental and sub-segmental level," in *Tenth Annual Conference of the International Speech Communication Association*.

Schuppler, B., Ernestus, M., Scharenborg, O., and Boves, L. (**2011**). "Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions," J. Phonetics **39**(1), 96–109.

Stevens, K. N. (**2002**). "Toward a model for lexical access based on acoustic landmarks and distinctive features," J. Acoust. Soc. Am. **111**(4), 1872–1891.

Story, B. H., and Bunton, K. (**2010**). "Relation of vocal tract shape, formant transitions and stop consonant identification," J. Speech Lang. Hear. Res. **53**, 1514–1528.

Yun, S., Choi, J. Y., and Shattuck-Hufnagel, S. (**2017**). "A landmark-based approach to transcribing systematic variation in the implementation of /t, d/ flapping in American English," J. Acoust. Soc. Am. **141**(5), 3583.