Acoustic Cues, Landmarks, and Distinctive Features: a Model of Human Speech Processing

Janet Slifka, Non-member

.

ABSTRACT

Four aspects of human speech processing are discussed along with their impact on the fundamental structure of a model of the human lexical access process (Stevens, 2002): (1) the lexical representation, (2) sensitivity observed in auditory processing, (3) multiple and graded activations of lexical candidates, and (4) contextual variation. The model assumes that the lexicon is represented in terms of basic units of sound contrast (distinctive features), and that non-homogeneous acoustic cues present in both coarse changes and finer details are used to estimate probabilities for the presence of underlying features. Acquired distributions of cue variation and associated dependencies are used to re-evaluate feature probabilities as context is extracted throughout the process. Existing feature modules, in general, correctly estimate features with a probability greater than 0.5 for 75-95% of their occurrences in read speech.

1. INTRODUCTION

A model aims to capture the fundamental principles of the process under study, in this case, speech processing by humans. Once established, adding complexities makes the model more realistic, although there are always trade-offs between tractability and accuracy. In this paper, we discuss four aspects of humanspeech processing and the manner in which these observations are incorporated into the Lexical Access from Features (LAFF) model, a model which has been under development by Stevens and colleagues for over ten years (e.g. Stevens, 2002; Slifka et al., 2004; Stevens, 2005). The LAFF model has two components: a theoretical framework based on studies of human performance and a software implementation that provides a platform for testing and refinement of the theory. The four aspects of human speech processing under consideration are: (1) the assumed form for the lexical representation, (2) the types of sensitivity observed in human auditory processing, (3) evidence for partial representations and graded activations, and (4) the time course for incorporation of contextual information.

These four observations determine the fundamental structure of the model and are outlined in Table 1. Each of the following sections details an aspect of the model structure: the representation of the lexicon in terms of distinctive features, detection and evaluation of acoustic cues to features, representation of feature estimates in a probabilistic format, and structures for re-evaluating feature probabilities based on the current known context. At present, development of the model focuses on the fundamental principles of acoustic processing for estimation of distinctive features and the use of context in refining these estimates. At this time, we are not considering the important role of syntactic and semantic constraints in this process, but are designing the model with flexibility to incorporate these additional complexities.

Table 1: *Overview of four aspects of human speech processing and their influence on the structure of the LAFF model.*

Human Processing	LAFF Model
Form of the phonological representation	Distinctive features
Sensitivity in auditory processing	Acoustic cues – coarse and fine analysis
Partial representations and graded activations	Probabilistic processing
Use of contextual information	Re-evaluating feature probabilities

2. PHONOLOGICAL REPRESENTATION

In automatic speech recognition (ASR) systems, the aim is to convert a speech signal into a sequence of words. While models of human lexical access also attempt to find the best match between the signal and a word sequence, the model is designed around the assumed mental representation of the lexicon. The listener's conversion of a continuous signal into a discrete sequence of words implies that the acoustic signal contains cues that allow the listener to perceive

Manuscript received on September 5, 2007.

The author is with Speech Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology,Cambridge, MA, USA 02139; E-mail: slifka@speech.mit.edu

contrasts. We assume, based upon a vast body of work in Linguistics and other fields that the lexicon is built upon the basic contrastive unit of the distinctive feature (Jakobson et al., 1952; Chomsky and Halle, 1968). This basic contrastive unit cannot be broken down any further. These features are binary, where changing the binary value of one feature in a word can potentially change it to a different word. The features that the LAFF model assumes are listed in Table 2. (Section 3 discusses the features as grouped into the landmark stage or as arising from finer acoustic detail.)

Table 2: *The inventory of features used in the LAFF model.*

Features from the landmark stage	**Features from finer acoustic detail**		
	vowel- or glide-related features	consonant-related features	
vowel consonant glide continuant sonorant	high low back tense spread glottis	strident lips tongue blade tongue body distributed anterior	lateral rhotic nasal voiced round

Each word is assumed to be represented in memory as sequences of segments where a segment is defined as a bundle of binary distinctive features (Stevens, 2002). The complete set of features is posited to be universal in language, and English uses an inventory of about 20 such features. Detection of 6 or 7 features is usually sufficient to identify a given segment. For example, the words "bill" and "pill" differ in one feature [voiced], which is related to the creation of a sound source from vibration of the vocal folds. The words "bit" and "beet" differ in another feature, [tense] which is related to the narrowness of the constriction in the oral tract during the vowel. The feature bundles associated with the contrasting sounds in these examples are listed in Table 3. In the remainder of the paper, the word "feature" will be used to refer to these binary distinctive features. This usage is in contrast to the common usage in ASR research of "feature" as any type of measure made on the acoustic signal (such as in the phrases "feature vector of mel-frequency cepstral coefficients" or "acoustic features for robust ASR.") In the LAFF model, the word "feature" is reserved for the abstract mental representation in the lexicon and the word "cue" is used to refer to measurements in the acoustic signal.

If circumstances such as noise, context, or speaking style lead to regions of the signal with inadequate information about a particular feature, other features

Table 3: *Example feature bundles for sound segments in English.*

/i/ (iy)	**/I/ (ih)**	**b**	**p**
vowel +	vowel +	consonant +	consonant +
high +	high +	continuant -	continuant -
low -	low -	sonorant -	sonorant -
back -	back -	lips +	lips +
tense +	tense -	voiced +	voiced -

in the segment are still likely to be adequately represented in the signal, i.e. acoustic cues to some of the features may be present while cues to other features may be degraded. For example, given that acoustic cues are measured in specific regions of the time-frequency space, transient or band-limited noise might corrupt a subset of the cues and leave others unaffected. Because the lexical representation is feature-based, the process of finding the best match between the signal and a word sequence has the flexibility to work from partially-specified feature bundles without the requirement to place a unique label on each bundle as a whole (such as a phone or phoneme label).

Each feature is associated with an acoustic and an articulatory representation. This representation is organized into two classes; (1) there is a defining articulatory and acoustic correlate that comes from relations among particular anatomical/acoustic/ perceptual attributes of speech sounds, based on what has been called "quantal theory" (Stevens, 1989); and (2) additional articulatory gestures are introduced in certain contexts of a feature to enhance its perceptual saliency (Stevens et al.,1986; Keyser and Stevens, in press). For example, for the feature [+nasal] the defining articulatory property is an opening of the velopharyngeal port in a particular range of areas, and the defining acoustic properties are the appearance of a nasal resonance in a particular range of frequencies and a concomitant flattening of the spectrum in the first formant range. The enhancing gestures for a feature are expected to depend on a variety of factors such as the range of sound contrasts in a given language and the phonetic and prosodic context in which the feature occurs. Typical examples for English are (1) the spreading of the glottis during a voiceless stop consonant closure and into the onset of an adjacent vowel, and (2) lip rounding in the production of [ ] (sh).

Knowledge of these enhancing gestures, together

with the defining gestures, and their acoustic correlates, is built into the model, and guides the acoustic analysis that leads to estimation of the features. In other words, the set of cues used to detect the presence of a specific feature depends on the articulatory actions associated with that feature and the expected variation in those actions based on context. For example, the cues to estimate the feature [high] are different from the cues to estimate the feature [rhotic]. The specification of these gestures and knowledge of articulatory-to-acoustic mappings provide a principled structure for extracting the acoustic cues.

The acoustic processing in the model has two general stages: (1) measurement of acoustic cues to features, and (2) estimation of the presence of features based on the cues. The range of challenges in executing these two aims includes fundamental questions such as how to extract acoustic measures that appropriately reflect the acoustic correlates (defining and enhancing), and how to assess the contribution of the cue values to a feature given the wide range of contextual variation. Section 3 discusses some aspects of the model structure that guide the measurement of acoustic cues to features, and Sections 4 and 5 discusses the representation of features in a probabilistic framework.

3. SENSITIVITY IN AUDITORY PROCESSING

A hallmark of human sensory systems is their marked sensitivity to abrupt changes. Abrupt acoustic changes during speech are created by specific actions of the articulators such as obstruction of the vocal tract, changing the sound source from vocal fold oscillation to noise, or changing the sound output path from the oral cavity to the nasal passages. In speech perception, humans are also known to be sensitive to a remarkably wide range of acoustic-phonetic detail that relates not only to the sequence of sound segments but also to aspects such as syllable structure, prosodic boundaries, turn-taking, and speaker indexical information. In the acoustic processing stage of the LAFF model, these two types of auditory sensitivity are reflected as two types of acoustic cues. Relatively coarse measures of energy patterns in frequency bands are used to detect instances of abruptness or maxima, where these instances are referred to as 'acoustic landmarks.' Distributed in the region around these landmarks are cues of the second type; cues which are particularly rich in information about the actions of the articulators that created the abruptness or local maxima.

The presence of a landmark indicates that the features for an underlying feature bundle (segment) should be measured. Landmarks are generally grouped into three basic classes based on the particular character of the abruptness or maxima: consonant landmarks (closure or release), vowel landmarks, and glide landmarks. (Stevens, 2002; Liu, 1995; Howitt,2000; Sun,1996)

The acoustic cues used in detecting landmarks are also used to specify the features [vowel], [glide], [consonant], [sonorant], and [continuant]. An example of detected landmarks in a simple sentence is given in Figure 1. Vertical bars mark locations of landmarks. In Figure 1a, at the landmark indicated by the arrow, the speaker releases a narrow constriction in the oral cavity and moves to a relatively open vocal tract configuration for the vowel with a sound source at the glottis. In Figure 1b, detected vowel landmarks mark a peak in low frequency energy.

Based on the feature set determined in the landmark stage ([vowel], [consonant], etc), cues measured in the vicinity of the landmark are used to specify the remaining features in the underlying feature bundle. For example, at a landmark associated with the feature [vowel], cues are measured to estimate dependent features such as [high], [low], and [back] but not features such as [strident], [voiced], [lips], [tongue blade], or [tongue body]. (See Table 2 for a division of dependent features.) Given that the model needs to detect roughly 6 or 7 features for a segment, one to three features are expected to be specified in the landmark stage (from relatively coarse acoustic cues), and two to four features are generally estimated in the second stage of finer acoustic analysis.

From the theoretical framework for basic and enhancing cues, from knowledge of articulatory-to-acoustic mappings, and from expected contextual dependencies, a set of measurable cues for implementation in the software model is specified where the cues are constrained to: (1) capture the relevant acoustic cue description (such as "spectral shape of the release burst"), (2) be appropriately normalized, and (3) make use of the entire frequency range for speech. Basic algorithms for estimation of energy within frequency bands, quantification of rate of change, and detection of local peaks (or dips) are the key components in the estimation of acoustic cues to features in the model.

In summary, the model assumes that instances of abrupt acoustic change and instances of local signal maxima are particularly rich in information about the actions of the articulators, and consequently are regions where acoustic cues to features are concentrated. In other words, the model does not assume that acoustic, phonological, and other information are uniformly encoded. The result is that the model does not use a frame-based approach with a uniform signal representation (such as MFCC and corresponding delta measures). The LAFF model processes the signal in a hierarchical manner where abruptnesses and peaks in coarse acoustic parameters guide subsequent processing of phonetic detail.

4. PARTIAL REPRESENTATIONS AND GRADED ACTIVATIONS

Early software implementations of aspects of the LAFF model used threshold-based methods to determine the presence or absence of each binary feature. Faced with a region of speech in which the acoustic cues are ambiguous, the model would still make a hard decision. Among the limitations with this approach are: (1) hard binary decisions on feature values limit flexibility in accessing the lexicon to determine the best match and (2) fixed thresholds limit the model's ability to capture the range of phonetic variation.

In addition, a range of current experimental evidence suggests that multiple lexical candidates are maintained during the human lexical access process. Each candidate is associated with a graded neural activation level where the activation is strengthened or inhibited as the lexical access process proceeds (e.g. Marslen-Wilson, 1987). Studies such as Spivey et al. (2005) and Allopenna et al. (1998) support continuous dynamic graded activation of multiple competing candidates during real-time spoken word recognition. In this view, the lexical access process is not the result of modular components cascading hard decisions forward.

Probabilistic models are the cornerstone of most speech processing systems as well as most cognitive models and are particularly suited to representing gradient information. The current implementation of the LAFF model assigns probability estimates to features on the assumption that listeners develop an experience-based knowledge of the distribution of cue values. Expected cue variation is part of the internal processing structure of the model that allows for more robust contact with the underlying features in the lexical representation. For example, in the process of assembling a cohort of word candidates from the lexicon, a feature with a weak probability may not cause a lexical item to be excluded and evidence from other non-acoustic sources could strengthen the overall probability of a lexical item.

5. CONTEXTUAL VARIATION

The observed phonetic variability that arises from context - e.g. surrounding consonants and vowels, syllable affiliation, prosody, social situation, and speaking style - raises the question of how humans recognize speech in the face of such variation. This large and long-standing question forms the basis for most, if not all, research on speech communication. In relation to a model of human speech processing, the question could be framed as: at what level(s) does the model account for such variability?

In most current ASR systems, contextual dependencies are typically captured by higher-order phone-based models such as tri-phones or quint-phones. These phone-based models can be limited in their ability to take full advantage of the range of contextual dependencies. For example, such models tend to under-utilize information from prosodic context, which can help to delineate utterance boundaries, detect stressed syllables, and interpret intonation patterns.

In some models of human processing, the variation is captured with the formulation of an exemplar-based lexicon, i.e. the lexicon stores exemplars of every experience of a spoken word as an essentially unanalyzed auditory token (e.g. Johnson, 1997). The model then statistically determines a structure of phonetic variation. This statement implies that variation is stored in the underlying representation and that the representation is updated every time we hear the word.

In the LAFF model, it is assumed every new experience of a spoken word does not alter the underlying lexical representation but rather has the potential to alter the principled process of cue selection, extraction, and weighting. Essentially, both approaches take into account the power of statistical representations in estimating the word sequence from the data in the signal. In the LAFF model, the probability of a feature is estimated from an acquired distribution for cue values where this acquired distribution is built from our past experiences of the cues to the feature in given contexts. (In practice, the model uses distributions based on training data.) In both formulations, new instances of spoken words contribute to our ability to process spoken language. The difference lies in where the influence is exerted - in the mental representation as an exemplar or as part of the process of principled cue extraction.

In the current implementation of the LAFF model, early stages of acoustic analysis in which only a limited context, if any, is available, may identify features with a low confidence level (weak probability). As additional information becomes available, whether it be information from sources such as features in the same segment, features in adjacent segments or in the same syllable, cohorts of words that are consistent with current feature estimates, position within a syllable, or proximity to prosodic boundary, the confidence with which a feature or a word can be estimated will increase. Essentially, relationships in the signal are re-evaluated as new sources of information become available. We are implementing a range of contextual re-evaluations that are expected to occur fairly often in normal speech and are likely to result in a better feature estimate. For example, the formant cues for stop consonant place of articulation can be more effectively evaluated if the feature [back] is known for the adjacent vowel (Suchato and Punyabukkana, 2005).

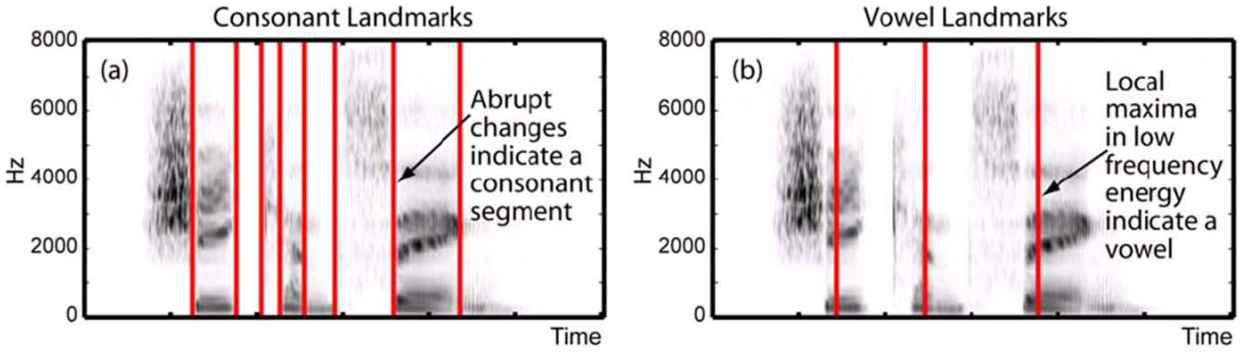

Fig.1: *Demonstration of acoustic landmarks for the utterance "She can sing." (a) consonant landmarks are marked by vertical lines (b) vowel landmarks are marked by vertical lines.*

6. PERFORMANCE OF THE SOFTWARE IMPLEMENTATION

Currently, the software implementation of the model has modules in varying states of completion for estimation of the features [vowel], [consonant], [glide], [sonorant], [continuant], [strident], [high], [low], [tense], [nasal], [voiced], and place features for stop consonants: [lips], [tongue blade], [tongue body]. This section is intended to briefly survey the type of performance results currently available in the model for feature estimation. In general, some of the components are more fully developed (e.g. [vowel], [consonant], [continuant], [sonorant], and stop place of articulation) and others are more preliminary in nature (e.g. [high], [low], [tense], [nasal], and [strident]). For evaluation purposes, features estimated with a probability greater than 0.5 are considered 'correct.'

A reformulation of consonant landmark detection into a probabilistic framework using the cue set from Liu (1995) detects discontinuities associated with the onset and offset of vocal fold vibration with 85% accuracy, discontinuities associated with sonorant consonants with 83% accuracy, and discontinuities associated with obstruent consonants (the burst release) with 87% accuracy. The data are from 24 speakers from the TIMIT database (Lamel et al., 1986) (3 speakers from each dialect region).

Irregular phonation in English serves both as a feature cue (such as [voiced] for voiceless stop consonants) and as a marker of prosodic structure. Automatic classification of tokens as instances of either regular phonation or irregular phonation based on four acoustic cues results in over 90% accuracy using support vector machines (Vapnik, 1995). Training and test data are from all speakers in 'dr1' and 'dr2' in the TIMIT database, where 114 of the speakers are used for training and the remaining 37 speakers are used for testing (Surana and Slifka, submitted).

For classification of stop consonant place of articulation: (1) stop bursts are classified with a greater than 90% accuracy; (2) conditioning on [voiced] and [back] in the adjacent vowel leads to a better classification accuracy in some contexts; and (3) for stops between two vowels, using cues from both vowels yields a classification accuracy of 95.5%. Burst spectrum cues contribute most effectively to classification, and formant transition cues are somewhat less effective (Suchato, 2004a; Suchato, 2004b).

The feature [tense] is correctly estimated in about 80% of the occurrences during read speech for two male speakers using a limited cue set of first formant (F1) slope and second formant (F2) slope (Slifka,2003). F1 is expected to decrease in [+tense] vowels as the articulators move to a very narrow constriction in the oral tract, and [-tense] (or lax) vowels in English are expected to show an offglide toward a neutral vocal tract (as measured in F2 slope).

Using only one cue, F1 minus F0 expressed in bark, as measured at the vowel landmark, the feature [high] is detected with 76% accuracy and [low] is detected with 78% accuracy from a database of 654 vowels from read speech for two male and two female speakers.

7. SUMMARY AND CONCLUSIONS

The LAFF model continues to evolve as new data are available on the human lexical access process, especially from the fields of linguistics and cognitive psychology, and as robust techniques are incorporated from the fields of statistical processing and machine learning. The core acoustic processing of feature cues is based on the guiding principles of defining and enhancing correlates to each feature and the relationship to contextual variation and language dependence. By placing the focus on achieving the 'best possible' performance based on acoustic cues, the model aims to provide a robust and flexible platform for future use of syntactic, semantic, and other higher level constraints and influences.

ACKNOWLEDGMENT

Supported in part by grant DC02978 from the National Institutes of Health. This work is in collabora-

tion with Ken Stevens and colleagues, and the author would like to thank Lisa Lavoie for helpful comments.

References

[1] Allopenna, P.D., Magnuson, J.S., & Tanenhaus, M.K. (1998) "Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models," Journal of Memory and Language, 38, 419-439.

[2] Chomsky, N. and Halle, M. (1968) *The Sound Pattern of English*, New York: Harper and Row.

[3] Howitt, A. (2000) Automatic syllable detection from vowel landmarks, Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA.

[4] Jakobson, R., Fant, G. and Halle, M. (1952) "Preliminaries to speech analysis: The distinctive features and their correlates," *Acoustics Laboratory Technical Report 13*, Massachusetts Institute of Technology, Cambridge, MA. Reprinted by MIT Press: Cambridge, MA, 1967.

[5] Johnson, K. (1997) "The auditory/perceptual basis for speech segmentation," *OSU Working Papers in Linguistics*, 50, 101-113.

[6] Keyser, S.J. and Stevens, K.N. (accepted) "Enhancement and overlap in the speech chain," *Language.*

[7] Lamel, L., Kassel, R., and Seneff, S. (1986) "Speech database development: Design and analysis of the acoustic-phonetic corpus," *Proc. DARPA Speech Recognition Workshop*, Report no. SAIC-86/1546.

[8] Liu, S.A. (1996), "Landmark detection for distinctive feature-based speech recognition," *Journal of the Acoustical Society of America*, 100 (5), 3417-3430.

[9] Marslen-Wilson, W.D. (1987) "Functional parallelism in spoken word-recognition," *Cognition*, 25, 71-102.

[10] Slifka, J., (2004) "Automatic detection of the features [high] and [low] in a landmark-based model of speech perception," *Journal of the Acoustical Society of America*, 115, 2428.

[11] Slifka, J., (2003) "Tense/lax vowel classification using dynamic spectral cues," *Proceedings of 15th International Conference of Phonetic Sciences*, Barcelona, Spain, 921-924.

[12] Slifka, J., Stevens, K.N., Manuel, S., and Shattuck-Hufnagel, S. (2004) "A landmark-based model of speech perception: history and recent developments," *Proc. of From Sound to Sense: Fifty+ Years of Discoveries in Speech Communication*, Cambridge, MA, C85-C90.

[13] Stevens, K.N. (2005) "Features in speech perception and lexical access," In *The Handbook of Speech Perception*, D. Pisoni and R. Remez (eds.), Blackwell Publishing: Oxford, UK, 125-155.

[14] Stevens, K.N. (2003), "Acoustic and perceptual evidence for universal phonological features," *Proceedings of 15th International Conference of Phonetic Sciences*, Barcelona, Spain, 33-38.

[15] Stevens, K.N. (2002) "Toward a model for lexical access based on acoustic landmarks and distinctive features," *Journal of the Acoustical Society of America*, 111, 1872-1891.

[16] Stevens, K.N. (1989) "On the quantal nature of speech ," *Journal of Phonetics*, 17, 3-45.

[17] Stevens, K.N, Keyser, S.J. and Kawasaki, H. (1986) "Towards a phonetic and phonological theory of redundant features," *In Invariance and Variability in Speech Processes*, J. Perkell and D. Klatt (eds.), Lawrence Erlbaum: Hillsdale, 426-449.

[18] Suchato, A. (2004a) "Classification of stop consonant place of articulation: Combining acoustic attributes," *Proc. of From Sound to Sense: Fifty+ Years of Discoveries in Speech Communication*, Cambridge, MA, C197-C202.

[19] Suchato, A. (2004b) "Classification of stop consonant place of articulation," *Journal of the Acoustical Society of America*, 115, 2629.

[20] Suchato, A. and Punyabukkana, P. (2005): "Factors in classification of stop consonant place of articulation", *INTERSPEECH-2005*, Lisbon, Portugal, 2969-2972.

[21] Sun, W. (1996) "Analysis and interpretation of glide characteristics in pursuit of an algorithm for recognition," M.S., Massachusetts Institute of Technology, Cambridge, MA.

[22] Surana, K. and Slifka, J. (in submission) "Towards a robust classification of regular and irregular phonation in normal, voiced speech."

[23] Vapnik, V. (1995) *The nature of statistical learning theory*, New York: Springer Verlag.

Janet Slifka was born in Ohio, U.S.A., in 1964. She received the B.S. and M.S. degrees in electrical engineering from the University of Dayton, Ohio, U.S.A., in 1987 and 1989, respectively. From 1985-1994, she was at Wright-Patterson AFB in the fields of satellite communications (1985-1987) and biocommunications (1987-1994). Following completion of her PhD in the Harvard-MIT Division of Health Science and Technology (2000), she spent time as a Fulbright Scholar in Portugal, and as an Acoustics Engineer for Bose Corporation. In 2002, she joined the Speech Communication Group at MIT as a Research Scientist. Dr. Slifka currently works for Eliza Corporation, MA, and teaches in the Boston area. Her research interests include speech respiration, acoustic cues to linguistic contrasts, and modes of vocal fold vibration.