ACOUSTIC CUE CONTRAST PATTERNS AND THEIR IMPLICATIONS

Thomas R. Sawallis
Florida Gulf Coast University, Ft. Myers, Florida, USA

ABSTRACT

Our implicit schema of a speech cue as consisting of paired acoustic distributions straddling a boundary presents both analytical and experimental problems. A better model can be based on the listeners' memory of the distribution of cues they have heard. This monomodal model alleviates the problems of the paired schema, and it explains and autonomously quantifies token value and abstract cue strength in revealing ways, particularly for speech perception and cross-language comparison, but also for phonology and other linguistic subfields.

1. THE ARCHETYPAL SCHEMA OF CUES

Speech cues are typically thought of as definable, measurable acoustical phenomena which help signal the identity of segmental phonemes. We expect that cue measurements should group in two distributions, associated with opposing feature values and separated by a clear valley representing the boundary between the features. Stop consonant duration exemplifies the archetype, with short voiced stops opposed to long voiceless stops. This cue schema, in some sense, inherits the characteristics we ascribe to features and segments, including the fundamentally structuralist nature of their opposition.

We understand a cue's location on the schema's scale as representing the strength of the cue in that token. A cue token in the tail of its distribution distal from the boundary is a strong cue for that feature, and a cue token near, or even over the boundary is a weak cue.

Each opposition is signaled by an array of several cues [2]. This redundant signaling is essential, since any weak cues are linked in trading relationships with the feature's other cues, which retain sufficient strength to preserve the intended percept.

2. THE MODALITY PROBLEM

A major weakness with the archetypal schema involves the "defective structure" resulting when the contrasting feature lacks the corresponding cue. Consider two known cues to stop consonant [voice] [2]. Consonant duration fits the archetype, opposing short voiced to long voiceless stops across a supposed phoneme boundary. We can adjust the strength of a token's percept by shortening or lengthening the closure of either voiced or voiceless stimuli, since both have a duration. However, we can adjust release burst+aspiration intensity only in voiceless stops, since voiced stops lack that segment to manipulate. Burst intensity can be termed a monomodal cue contrast, and compared with the archetypally bi- (and potentially multi-) modal contrast of consonant duration. The monomodal pattern poses a problem because it lacks an opposing distribution to prompt a language learner to infer a boundary. Yet studies of token goodness and of prototypes [5, 6] lead us to expect listeners will treat "outliers" differently, despite the absence of the boundary.

3. CROSS-LANGUAGE CUE DIFFERENCES

Further difficulties arise from any attempt to compare cue structures across languages.

3.1. In Single Cue

The "same" cue may differ across languages in several ways, all detectable by appropriate surveys.

The simplest difference is that the cue distributions may not have the same location. For instance, as surveyed by Lisker & Abramson [3], the lead VOT ranges for /g/ in Hungarian and Marathi do not overlap. (Figure 1)

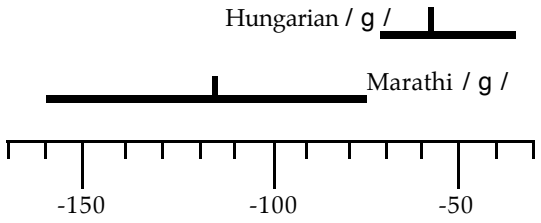

Figure 1. VOT distribution differences.

Under the paired cue schema, one, both, or neither of the distributions could be displaced. Among the possible combinations, that means the distributions in a second language could be closer or further apart, and they could be displaced in either direction. The cue distributions themselves may also differ in spread (greater, smaller, or the same range or variance) or in shape (skewness or kurtosis),[1] which adds another set of combinatorial possibilities.

These potential differences create problems for comparison of paired cue schemas across languages. If the configurations in two languages, A and B, were identical, but shifted, tokens occuring between the boundaries would cue one feature value in language A and the other in B. Consider the problem if instead, the variances were higher in B, so that the distance between the outside tails was greater, but the distance between the inside tails lessened to the point of overlap. The cue might be more informative in "strong" tokens, i.e., those in the more distant outside tails, while the cue's overall reliability suffered because of the overlap. A similar difficulty arises if the distribution means are more widely spread in B, with the boundary location and distribution size and shape the same as in A. Is the cue stronger by the separation, or does such a separation develop to compensate for the cue's weakness? Certainly, in L2 speech perception, cue tokens from A would seem weak to B speakers, and B tokens might seem exagerated to A speakers. These three comparative possibilities between languages A and B are illustrated in Figure 2.

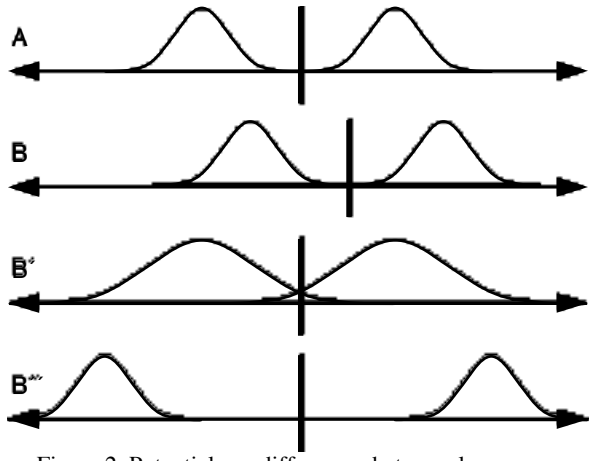

Figure 2. Potential cue differences between languages.

3.2. In Segmental Phonology

The phonological systems in which the cue is embedded differ across languages, and that may, directly or indirectly, influence the cue strength. This can occur in several ways.

English, like many languages, contrasts stops in 2 series at 3 places, yielding (/p, t, k/) vs. (/b, d, g/). Some languages with these series and places have gaps: Arabic lacks /p/ and Thai lacks /g/. Yet such gaps in the cue schema do not seem to disrupt the perception of the "unopposed" segment, Arabic /b/ or Thai /k/ [4]. This situation complicates the monomodality problem noted above since the cues are both monomodal in the unopposed segment and bimodal elsewhere in the series.

Consider next Lisker & Abramson's finding [3] that languages with three initial stop series oppose voiced, plain voiceless, and voiceless aspirated, while languages with two series oppose plain voiceless to one of the others, with the main cue for the three series being the lead, short lag, and long lag VOT ranges. Initially, this observation demands extending the paired cue schema to a trio and two boundaries. One difficulty, though is explaining what becomes of the unused cues in the two series languages. Generally, the two-series languages keep to their ranges. In English, however, the "gap" is appropriated: both lead and short lag VOT occur for /b/. Another difficulty: Does the larger phonemic decision space of the 3 series languages mean the cues are intrinsically more important due to reduction in overall redundancy? Does this answer hold in comparing entire languages with small (e.g., Hawaiian) and large (e.g., Georgian) phoneme inventories?

Syllable patterning could also affect the phonemic background of cues. Some languages (e.g., Mandarin) are restricted to little more than a CV syllabic framework, while others (e.g., English) allow diphthongs & glides in the nucleus and dense consonant clusters in onset and coda. Do listeners use cues to initial /t/ differently if they also have be able to recognize initial /str/?

3.3. In Suprasegmental Perceptual Skills

Cross language suprasegmental processing differences (at least) may also pose problems for interpreting cue strength according to the archetypal schema. Imagine two languages, T and L, with identical segment inventories and cue structures, where T is a tone language and L has phonemic vowel length. The Lisker List of stop voicing cues [2] allows us to expect both duration and f0-based cues in both T and L. We can reasonably expect the perceptual skills and habits of mature T and L speakers to differ with regard to f0 and duration, and to affect their use of stop voicing cues. Yet the archetypal interpretation of cue strength as distance from the boundary could not represent such a processing difference.

4. MAJOR RESEARCH DIFFICULTIES

This discussion suggests that well placed questions can reveal uncertainty regarding the linguistics and psychology of speech cues. These uncertainties have serious ramifications for the use and interpretation of some fairly common research designs.

The Lisker List [2] catalogs 16 cues with known effects on intervocalic stop voicing in English, and it is not a closed list. Some are mono- and some bi- or multimodal, but all are potentially used in other languages with other location and distribution settings than in English. Those setting differences would affect any putative boundary location, and also the patterns of the intercue trading so important for robust perception. The cues are then embedded in phonological structures which differ across languages and which may well favor different perceptual strategies.

We must conclude that English (or English-based synthetic) stop consonant stimuli have too many uncontrolled variables to be used for basic (i.e., non-contrastive) speech perception research with subjects other than English speakers. More broadly, we must apparently conclude that it is unsafe to generalize from speech research done with stimuli drawn from other than the subjects' native language.

The underlying tenet is this: There are no such things as language-neutral speech stimuli, because there are no such things as language-neutral listeners.

The repercussions go yet further. The critique above implies that cue strength has two aspects: token value determined by the location of the cue token in its distribution, and a more abstract strength applicable to the cue regardless of token value. The perceptual contribution of a cue token is determined jointly by the two factors.

Now recall that many speech perception stimuli are created from a single (natural or synthetic) base, with one or two cue experimental parameters varied through some range by digital editing (or in earlier times, tape splicing). The non-experimental cues are unlikely to have been checked as having been set to a central (for the language) token value. Indeed, one cannot even check, for instance, the three most important cues, because the literature appears to lack any method for determining abstract cue importance. Thus, it is impossible to state, for stimulus series so constructed, whether there are important cues set to extreme values skewing the experimental results through trading relationships.

5. RECASTING THE CUE

The archetypal "paired cue plus boundary" schema is the simplest melding of the phonetic fact of cue variation and the phonological fact of contrast, so its attractiveness is natural. But the critique above suggests it is of problematic applicability in studying the fundamental perceptual nature of cues, especially in the sort of cross-language comparisons which are so useful in other branches of linguistics.

Yet the critique also suggests desiderata for an improved model of acoustic speech cues. It should provide: principles for comparison of mono-and bi-modal cues; principles for

comparison of cues across languages; testable hypotheses regarding phonetic perception of contrasting segments; and separate estimates of token strength and general cue importance.

5.1. Memory

In fact, a model fulfilling those desiderata can be built on a simple claim: Listeners remember the cues they hear and the context in which they are heard. More specifically, listeners have a long term memory representation of the central tendency and level of precision – effectively, the distribution[2] – of the cues they use in speech perception. These representations both function with and retain information about a rich variety of factors associated with the context in which the cues are used, including at least feature and phoneme identity and phonological context, and possibly speech rate, dialect, and other factors.

The listener's representation of the cue distribution is more than a memory; it is a distillation of the acceptable cue utterances in the listener's community. It describes the community's conventional target for that cue, and the listener's expectation in future utterances. The interpretation of cue memory as cue expectation is crucial, since it creates analogies across cues.

5.2. Tokens and Analogies

Assume (rhetorically and falsely) that all cues are normally distributed on acoustic (rather than psychoacoustic) dimensions. The level of expectation of a cue which occurs at its distribution's mean is equal to the expectation of any other cue at its mean, regardless of (mono- or multi-) modality, feature, phoneme, context, and even language. In the example in Fig. 1, a VOT for /g/ of -58 ms in Hungarian is perceptually equivalent to one of -116 ms in Marathi, since they are equally expected by native listeners. They are also equivalent, for instance, to the 2655 Hz surveyed mean of /e/ F2 by female speakers for Southern California English [1].

Extending the analogy, we can assert equivalent expectation of any cues which occur at other specified locations in the distributions, for instance, at 2 standard deviations (s.d.) weaker (relative to some contrast) than their means. Thus, a Southern California female token of /e/ with F2 at 3029 Hz contrasts with /i/ the same as a Marathi /g/ with a VOT of about 90 ms[3] contrasts with /k/. By further analogy, we can assert that any standardized modifications of cues are equivalent: a 3 s.d. weakening of Hungarian VOT, Marathi VOT, and English F2 are equivalent, and 3 s.d. weakenings in Marathi VOTs of 150 ms and 80 ms are equivalent.

5.3. General Cue Weight

Expectation is thus the measure of the strength of cue tokens, and can be used to design equivalent modifications across cues. Those equivalent modifications can be used to test perceptual effect, and thereby to measure the abstract importance of the cue tested.

Assume one measured natural corpuses of intervocalic /g/ each in languages A and B has been tested on native speakers and both are misperceived as /k/ at a rate of 2%. Lengthen closure durations by a linguistically equivalent (but probably acoustically different) 3 s.d. in each corpus, and retest. If /k/ percepts increase to 4% in A and 6% in B, then closure duration is a more important cue for /g/ in B than in A. Start over with the unmodified A corpus, attenuate closure voicing amplitudes by 3 s.d., and retest. If /k/ percepts increase to 7%, then in A, voicing amplitude is a more important cue for /g/ than is closure duration.

A preliminary study [7] similar to this last example has been done on four cues to intervocalic /t-d/ in French. Results showed that native French speakers use closure duration more in perception of /t/ than /d/, and suggested that non-natives use cues differently than do natives. The study also showed how abstract cue strengths are autonomously quantifiable, and can be compared validly across conditions, languages, etc.

6. IMPROVED CUE MODEL

6.1. Overview

A cue model explicitly based on listeners' memory offers two caveats and a number of advantages over the implicit schema of paired, universally specified distributions straddling a boundary.

The caveats: 1) There are no stimuli which are processed as equivalent speech tokens by native speakers of different languages, since the native language experience is a crucial factor in speech perception. 2) Generating perceptual stimuli by iteratively modifying a base will fix non-experimental cues at uncontrolled values and may modify their memory representation, thereby skewing the experimental results.

The advantages: 1) Though memory itself is not observable, a close approximation is accessible via normal phonetic survey methods. 2) Comparisons based on distributional analogies can be made across cues, features, phonemes, contexts, dialects, languages, etc. 3) "Self-calibration" by distributional criteria quantifies token values. 4) Mastery of token values allows separate investigation and autonomous quantification of abstract cue weight. 5) Phoneme boundaries and cue trading ratios are epiphenomena, calculable from autonomous token and cue factors. 6) Since mono- and bimodals are treated alike, no problem is caused by monomodals' failure to motivate a boundary.

6.2. Implications

The monomodal memory model of cues implies that the redundant encoding of the "same" contrasts and segments may in fact be managed and perceived differently across languages. This has ramifications for explanations of allophonic and phonotactic patterning. The model could also be useful in L2 speech pedagogy, since it claims that a cue's importance and its target may be separable locuses of L1 interference.

The model also has potential implications for other sub-fields of linguistics. For instance, in language acquisition, the loss of infants' ability to distinguish certain non-native speech sounds could be explained by initialization of the memory representation of native cues, or by redirecting attention to the cues important in the native language. Also, in historical linguistics, sound change per se clearly affects acoustic cue distributions, and might depend on certain cue strength conditions, but phonemic restructuring — mergers and splits — might be accomplished by cue strength shifts alone, with no acoustic change needed.

NOTES

1. Henceforth, for expository simplicity', this paper makes the false assumption that cue distributions approximate to normal curves.
2. The shape of the distribution and the accuracy of its representation are subject to investigation.
3. This is a guess. Standard deviations were not reported in [3]

REFERENCES

[1] Hagiwara, R. 1997. Dialect variation and formant frequency: The American English vowels revisited. Journal of the Acoustical Society of America, 102, 655-658.

[2] Lisker, L. 1986. "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. Language and Speech, 29, 3-11.

[3] Lisker, L. & Abramson, A.S. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. Word, 20, 384-422.

[4] Lisker, L. & Abramson, A.S. 1970. The voicing dimension: Some experiments in comparative phonetics. In Proceedings of the Sixth International Congress of Phonetic Sciences, 1967 (pp. 563-567). Prague: Academia.

[5] Miller, J.L. 1994. On the internal structure of phonetic categories: A progress report. Cognition, 50, 271-285

[6] Rosch, E. 1977. Principles of categorization. In Rosch, E, & Lloyd, B. B. (eds.), Cognition and Categorization (pp. 27-48). Hillsdale, NJ: Lawrence Erlbaum Associates.

[7] Sawallis, T.R. 1996. An autonomous system for quantifying the perceptual use of acoustic speech cues: Voicing in intervocalic /t-d/ in French. Unpublished doctoral dissertation. University of Florida, Gainesville, FL.