AUDITORY FEATURES FOR HUMAN COMMUNICATION OF STOP CONSONANTS UNDER FULL-BAND AND LOW-PASS CONDITIONS

Eduardo Sá Marta, Luis Vieira de Sá
email: EMARTA@CO.IT.PT
Dep. Engenharia Electrotécnica, FCTUC (Universidade de Coimbra)
Instituto de Telecomunicações - Pólo de Coimbra

ABSTRACT

A set of auditorily-formulated features for PLACE discrimination in stop consonants, uncovered in extensive experiments with natural and edited sounds, are now being modeled using fuzzy logic and being applied to large databases of monosyllabic and spelled letters speech sounds, in various languages, in full-band and low-pass conditions. The rationale is that any valid model of human communication should replicate the human listener "feats" of very good (albeit not perfect) discrimination of stop PLACE even from speakers of different languages, and of "graceful degradation" when faced with markedly low-pass filtered sounds (e.g., telephone-like).

This paper reports mainly about fuzzy-logical models, expressing known auditory phenomena, that evaluate the high-frequency content of the *burst+aspiration* segment in stop consonants, and provides a powerful cue for discrimination of DENTAL consonants. This evaluation is robust to mild variations of the frequency response curve such as those caused by different recording microphones.

These cues are absent from markedly low-pass filtered sounds, however. Other cues, which survive in low-pass sounds (reported in previous papers), are also discussed.

1 - INTRODUCTION

1.1 – High-frequency content of the *burst+aspiration* segment as a cue for the DENTAL category

It has long been acknowledged that DENTAL stop consonants (in consonant-vowel syllables) tend to show a *burst+aspiration* segment which is very rich in high-frequency energy, as compared to other stop consonants [1]. But it has also been intensely debated whether this high-frequency content is a 1st-order cue for the DENTAL category, or rather a 2nd-order cue - one which might readily be overruled (in listener's judgements of perceived PLACE) by cues pertaining to formant transitions [2].

The question remains of establishing one or more auditorily-formulated metrics for "*burst+aspiration* high-frequency content" which show the type of performance evident in human listeners, namely the almost complete independence relative to mild variations in the frequency-response curve.

The merit of any proposed metric should be tested with sounds uttered by large numbers of speakers, which has not been the case with most past studies. Since it is also known that stop PLACE discrimination is done at very high levels of accuracy [3] across languages (that is, by listeners native of a language different from that of the speakers), a more stringent – and entirely called for – test is that of evaluation with (large numbers of) speakers native of different languages. That is, a single model/metric should be applied unmodified to several languages.

1.2 – Assumptions about the role of cues in human phoneme communication

Our assumptions, and the reasoning supporting them, are briefly given in this section.

As it has been underscored by many authors the mechanisms underlying human speech communication must have properties that provide some sort of "shield" against a number of variabilities in the speech communication channel: variations in the frequency-response curve, or noise characteristics of the communication channel are obvious detrimental factors. Heavy noise or drastic filtering may even cause "missing data". These problems are already present in simple communication tasks, such as the one of communicating the DENTAL vs. LABIAL distinction in stop consonants (we will henceforth refer to this task).

One plausible solution is the use of multiple cues, all being orthogonal to between-categories boundaries, for each phonemic distinction. There exist, among the cues, trade-off relations that may extend to the point of alternativity. Acoustic degradation is likely to attenuate some cues, whereas others survive. If the main effect of signal degradation is to attenuate the existing cues (not falsely creating instances of some cues), the end result of extreme degradation will be "cue squashing". Even if the listener is then unable to issue a reliable classification judgment, he will be able to tag the segment as "devoid of discriminatory information" and resort to higher-levels of message recovery (e.g. grammar, semantics).

Cues are not equipotent. Cues acknowledged by listeners (and speakers) as more reliable will be weighed more than other less reliable cues. This makes it conceivable that, in order to emit a valuable cue for the intended category, the speaker may indulge (because of articulatory constraints) the emission of a conflicting, but less weighed cue.

Thus, metrics for the 1st-order features should show high scores only for exemplars of the correct (associated to the feature) category. In contrast, it is admissable that some (not common, hopefully) exemplars of a particular category will show significant scores for some 2nd-order features which are "wrong", that is, "point" to another category; if 1st-order "correct" features are strongly present, they will overrule the "wrong" 2nd-order features.

In speech production acquisition by a child, once the vicinity of gross articulatory correctness is attained, auditory feedback may well be the only "evolutionary force". Since this auditory feedback incorporates trading relations between cues, the new speaker may "rest satisfied" when he/she learns to emit sufficient cues with substantial intensity. Different speakers may end up with very different mixtures of the available "palette" of cues.

From several experiments and results, we are convinced that any cue related to "*burst+aspiration* high-frequency content" must be a 2nd-order cue for the DENTAL category.

2 - FIRST AUDITORY METRIC FOR HIGH-FREQUENCY CONTENT IN THE *BURST+ ASPIRATION* SEGMENT

The reasoning behind this metric is that it must exhibit (as human listeners do) independence relative to mild variations of the frequency-response curve. One simple way to accomplish this is to take the difference in "high-frequency content" at roughly the same frequency range, at two different points in time – any variation will affect the two evaluations equally, and will therefore be cancelled by the difference. The two points in time should have a consistent definition and be reliably discernable by the listeners; the vowel onset clearly provides the required differentiation. One such metric has been proposed with some success [4] in the discrimination of DENTAL versus LABIAL stop consonants, but little – if any – tribute has been paid to auditory considerations in its definition; also, it has been tested with just a few speakers. The metric in [4] uses 2 arbitrarily-defined frequencies for the evaluations, whereas it is to be expected that human listeners' evaluations should be frequency-shift-independent to some degree; that is, the same pattern of change, whether occurring at a particular frequency or at another frequency (say, 20% lower, or higher) should have similar perceptual consequences.

The metric we propose aims to express the degree to which a listener may detect that the segment prior to vowel release (that is, the *burst+aspiration* segment) has a considerably stronger high-frequency content then the ensuing vowel. We assumed that this comparison is made through integrative mechanisms that have low resolution in frequency and in time, and that the vowel onset is used to differentiate a "BEFORE" comparison term and an "AFTER" term. Both terms were evaluated using 1.55KHz rectangular-bandwidth integration over input vectors which are FFT spectra calculated, with a Hamming window, over frames of 11.6ms, with a 3-ms frame advance. Since loudness growth of a bandpass signal is most rapid in the first few tens of milliseconds, we averaged energy over 30 ms (within 100 ms after vowel onset) and then took the maximum of these averages to represent the AFTER term; for the BEFORE term, we averaged over 15 ms since 30 ms would tend to under-represent short *burst+aspiration* segments, particularly in voiced stop consonants. Maxima were taken to be a rough representation of the acoustic events with the strongest impact on the auditory system.

We implemented a frequency-sliding evaluation : the existence of a drop in high-frequency content (from the BEFORE term to the AFTER term) is investigated for frequencies (corresponding to the lower-frequency limit F.BEFORE in the 1.55KHz integration interval of the BEFORE comparison term) from 4040Hz to 6190Hz, in 86-Hz steps. The AFTER comparison term extends from F.BEFORE-1.2KHz to the maximum frequency of the input vectors.

Such a drop may not perceptible if the BEFORE term is of too low energy. We expressed this through a intersective fuzzy factor whose physical parameter is the energy of the BEFORE term. Also, very intense low-frequency energy during the BEFORE term might diminish the perceptual impact of its high-frequency content. This called for a second intersective factor, connected to the absolute value of the low-frequency energy, and its relation to the high-frequency energy (all quantities were evaluated through maxima collection, for simplicity).

The time window for the BEFORE component was extended, on the face of results with sounds with very short *burst+aspiration* segments, to up to 9 ms into the vowel.

Automated optimization of this model is problematic: since this is just one of several contributions to DENTAL perception, it is not true that all DENTAL exemplars should exhibit high scores in this metric. On the other hand, since this is presumed to be a 2nd-order cue for DENTAL, and thus can be overruled by 1st-order cues for other categories (e.g., LABIAL) it is admissible that some LABIAL exemplars could elicit medium to high scores in the metric.

We refined the metric through inspection of its results in 5 sets of sounds: an in-house research database of /ti/ and /pi/ sounds from 33 Portuguese speakers (representative of Portuguese unvoiced stops), the letters "T" and "P" from the first set (30 speakers, 120 sounds) of the Oregon Graduate Institute ISOLET Database (representative of U.S. English unvoiced stops), the letters "D" and "B" from the first set of ISOLET (representative of U.S. English voiced stops), the letters "T" and "P" from the first 50 speakers of the Bavarian Archive for Speech Signals PHONDATA1 Database (representative of German unvoiced stops), and the letters "D" and "B" from the same 50 speakers (representative of German voiced stops). It is to be emphasized that the same model was used throughout, with no adaptation whatsoever, and that the different sets have obviously used different microphones, as well as recording conditions.

Refinement through inspection, with the added burden of not having entirely definite targets, is extremely time-consuming, and we have been able to explore only a few parameters (e.g., different frequency and time ranges were not explored). However, satisfactory results were obtained almost from the outset. The results obtained (for a model that elicited satisfactory results in all the sets considered) for the PHONDATA1 (1st 50 speakers) letters "T" and "P" are presented below in histogram form (the metric values, in the fuzzy range 0.0-1.0, correspond to the horizontal axis) :

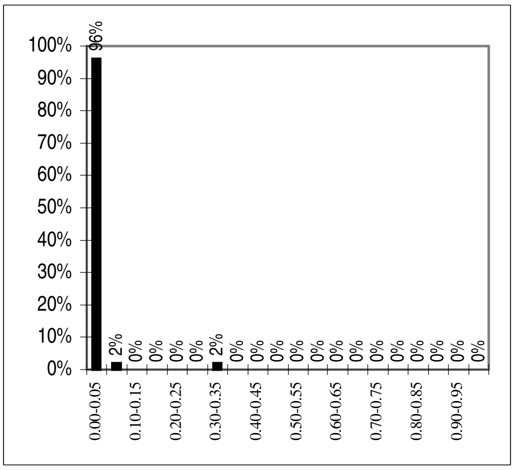

Figure 1 – Histogram for the 1st auditory metric elicited by 50 German "P" sounds

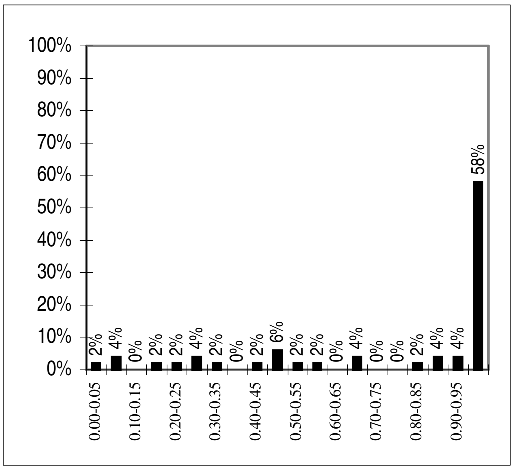

Figure 2 – Histogram for the 1st auditory metric elicited by 50 German "T" sounds

Used alone, this metric achieves 98% correct discrimination (using any threshold value between 0.01 and 0.025). However, this result is not especially important in the cadre of explaining human communication of this phonemic discrimination. First, since this metric is being proposed as just a contribution to the perception of the DENTAL category, the considerable spread of the "T" histogram is entirely to be expected: a good number of "T" exemplars (68% of "T" exemplars are above the 0.75 mark) show very high values for the metric (which means that in those exemplars the metric contributed substantially to the perception of the intended phonemic category) but many other "T" exemplars show only medium, or even low, values for the metric (which means that in those exemplars other cues were primarily responsible for the perception of the DENTAL category).

The almost complete squashing of the "P" histogram towards near zero metric values is in good accordance with the hypothesis that this metric works for the perception of the DENTAL category and against the perception of the LABIAL category. However, the single sound responsible for the 2% mark at 0.30-0.35 requires explanation. This single sound is jnedP (PHONDATA label) and inspection shows that this sound possesses a markedly ascending transition of the 2nd formant. Such a transition, particularly with a strong aspiration of the 2nd formant, is widely regarded as a powerful cue for the LABIAL category. Thus this is likely a case in which a strong 1st-order cue for LABIAL was emitted and (as a consequence of the energy "invested" in the onset of burst/aspiration) emission of a contrary, but 2nd-order, cue was indulged by the speaker.

3 - SECOND AUDITORY METRIC FOR HIGH-FREQUENCY CONTENT IN THE *BURST+ ASPIRATION* SEGMENT

One alternative way to achieve independence relative to mild variations of the frequency-response curve is to define a metric that is differential along the frequency axis. This metric evaluates the degree of existence of (local in frequency) upward inflections in the spectrum, occurring at "high" frequencies during the *burst+aspiration* segment. Since the inflection is evaluated along a short frequencial span, changes in the spectral slope do not appreciably disturb this evaluation. On the other hand, the mechanism of lateral inhibition, prevalent in the auditory system, makes this a highly plausible solution.

This metric is considerably more complex than the previous one. The core of the metric is a raw inflection, which corresponds, in each sound frame, to the dB increase from frequency **f** to frequency **f**+250Hz. A series of fuzzy intersective factors are used to express auditory mechanisms that tend to decrease the auditory representation of such inflections: lateral suppression from strong lower-frequency components, adaptation, backward-masking.

It is assumed that the perceptual impact of any spectral inflection depends on its duration. However, the form of the dependence is not known, although substantial growth is more likely to exist during the first few milliseconds, and be less important afterwards. This lack of knowledge was obviated by evaluating an averaging measure over 9 ms, over 12 ms, and over 15 ms. For each time duration, an exclusively-DENTAL range was estimated and this allowed fuzzification of the measures for each of the three durations, followed by fuzzy union of the three fuzzified measures.

The first and second metrics were subject to fuzzy union. The histograms corresponding to this union are shown below.

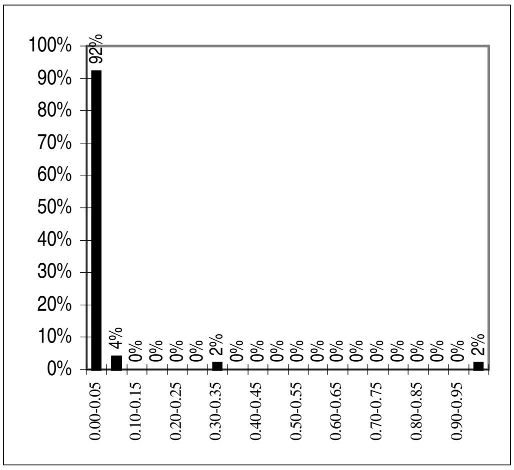

Figure 3 – Histogram for the fuzzy union of the 1st and 2nd auditory metrics elicited by 50 German "P" sounds

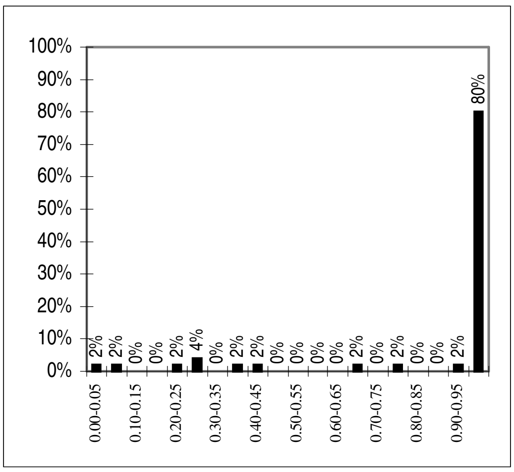

Figure 4 – Histogram for the fuzzy union of the 1st and 2nd auditory metrics elicited by 50 German "T" sounds

As to the “T” sounds, it is readily apparent that the union of cues tends to push all the exemplars towards high values. The “P” sounds remain solidly at very low values, with the sole exception of one sound, which is kprdP (PHONDATA label); again, this sound shows a markedly ascending F2 transition, which may overrule the 2nd-order DENTAL cue.

Results for other sets are also supportive of the hypothesis that these are reasonable formulations for a 2nd-order DENTAL cue. For instance, for the “B” and “D” letters in ISOLET1:

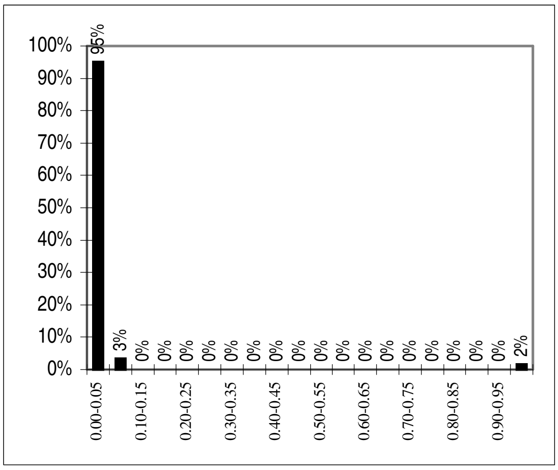

Figure 5 – Histogram for the fuzzy union of the 1st and 2nd auditory metrics elicited by 60 U.S. English “B” sounds

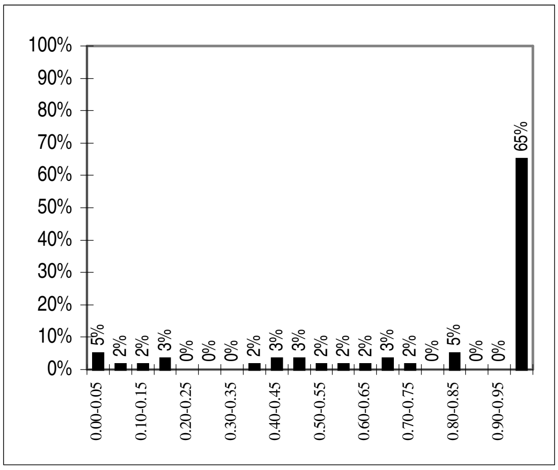

Figure 6 – Histogram for the fuzzy union of the 1st and 2nd auditory metrics elicited by 60 U.S. English “D” sounds

4 – CUES AVAILABLE UNDER LOW-PASS CONDITIONS

For severely low-pass sounds, such as in telephony, it is a well-known fact that the human discrimination between “P” and “T” letters, for instance, is considerably degraded, but remains high.

To explain these facts, it is necessary (in the cadre of the assumption of cues orthogonal to the between-classes boundary) to (i) define auditorily-inspired models for cues that are present in full-band sounds but not in telephonic-filtered sounds and (ii) models for cues that survive in telephonic-filtered sounds and can explain the observed level of human listeners’ performance.

The cues reported in sections 2 and 3 are precisely as required in (i) and may explain the much better discrimination observed for full-band sounds. Auditorily-inspired models for cues that survive in low-pass signals have been previously reported by the present authors. These models are for “ascending sequence cells” as a cue for LABIAL [5] and for “level-tolerant neurons” as a cue for DENTAL [6].

It has been shown that these low-pass-surviving cues can discriminate U.S. English “B” versus “D” sounds low-pass filtered at 3.5KHz at less than 7% errors.

Other cues have been already identified, but remain to be modeled: for instance, it has been recognized that sequence cells responding to approximately equal-frequency sequences may provide yet another cue for the DENTAL category.

ACKNOWLEDGEMENTS

The research reported in this paper has been conducted under the Research and Development Contract Praxis 2/2.1/TIT/1558/95 of the PRAXIS XXI Program of the Junta Nacional de Investigação Científica e Tecnológica.

REFERENCES

[1] - Stevens, K.N., and Blumstein, S. E. - "Invariant cues for place of articulation in stop consonants", J. Acoust. Soc. Am. 64, 1978, 1358-1368

[2] – Sussman, H.M., McCaffrey, H.A.; and Matthews, S.A. - "An investigation of locus equations as a source of relational invariance for stop place of articulation", J. Acoust. Soc. Am. 90, 1991, 1309-1325

[3] - Anna Marie Schmidt, "Cross-language identification of consonants. Pat 1. Korean perception of English”, J. Acoust. Soc. Am. 99 (5), May 1996, 3201-3211

[4] – Lahiri, A., Gewirth, L. and Blumstein, S.E., "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants”, J. Acoust. Soc. Am. 76, 1984,

[5] - Eduardo Sá Marta, Luis Vieira de Sá - "Impact of ascending sequence AI (auditory primary cortex) cells on stop consonant perception" - Proceedings of the *5th European Conference on Speech Communication and Technology Eurospeech 97* - Rhodes, Greece, 1997

[6] - Eduardo Sá Marta, Luis Vieira de Sá - "Auditory cells with frequency resolution sharper than critical bands play a role in stop consonant perception: evidence from cross-language recognition experiments" - Proceedings of the *NATO Advanced Study Institute on Computational Hearing,* Il Ciocco, Italy, 1998