RELATIVE CONTRIBUTIONS OF NOISE BURST AND VOCALIC TRANSITIONS TO THE PERCEPTUAL IDENTIFICATION OF STOP CONSONANTS

Adrien Neagu and Gérard Bailly
Institut de la Communication Parlée
46, av. Félix Viallet 38031 Grenoble CEDEX FRANCE
e-mail: (neagu,bailly)@icp.grenet.fr

ABSTRACT

A set of three perceptual experiments is described. These experiments were designed to provide identification scores on CV sequences for French. Original stimuli were augmented with acoustic "monsters" where burst were excised or replaced. The first identification task shows that information carried by vocalic transitions can be overwritten by burst information. The importance of this phenomenon is inversely proportional to vowel aperture. The second experiment shows that these results are almost insensitive to relative amplitudes between the burst and the vowel. In the third experiment we manipulated the voice onset time (VOT) of the monsters using high quality analysis-resynthesis. Stimuli with a very short VOT were perceived as bilabials but VOT manipulation did not affect the /t/-/k/ confusions. These experiments claim for a dynamic model of stop identification where burst and vocalic transitions both contribute and compete to the phonetic decision.

1. INTRODUCTION

The quest for acoustic correlates for the identification of stop consonants focus a large effort of speech research. Due to their strong coarticulation with adjacent segments it is difficult to point out simple or even relative invariance. The perceptual experiments described here have two main goals: (a) creating perceptual illusions with distorted but still highly natural stimuli to provide insights in human perception; (b) providing acoustic-to-phonetic decoders using large training database with stimuli they have no chance to observe and that could guide them towards the most informative part of the signals.

2. STATE OF THE ART

Many experiments with distorted stop consonant-vowel syllables aimed at finding acoustic correlates for the place of articulation. Two complementary attempts are the voiced transition-based metrics [10] as opposed to release burst-based ones [2]. Blumstein and Stevens [1] attempt to capture simple invariance through spectral integration. On the other hand, others [7, 6] claim that dynamic spectral changes from burst release into formant onset better capture relative invariance. Meanwhile, the way redundant cues encoding place of articulation in natural speech can conspire or overwrite each other was less studied [5, 11]. More recently, Smits et al [9] conducted a perceptual experiment similar to our Experiment I with 6 plosives and 4 vowels of Dutch. The present paper confirms their results but for the full French vocalic system (10 oral vowels). Furthermore, the importance of global characteristics of the burst such as relative amplitude and timing relative to the vocalic substrate is examined.

3. PERCEPTUAL EXPERIMENTS

3.1. Experiment I

Stimuli :

The original stimuli for the tests are 30 stressed CVs (/p/,/t/,/k/ * 10 vowels) embedded in nonsense words. This corpus was recorded at 16kHz from one male speaker. We extracted (zero cross cut) for each item two speech segments: (a) the noise-like segment (N-seg) up to the the first glottal period and (b) the voiced segment (V-seg) from this first glottal period to the end of the syllable. Acoustic monsters are built up by carefully padding N-seg and V-seg segments in order to avoid clicks generation. For example, /pa/ V-seg received /ta/ and /ka/ N-segs. We obtain 60 items built from conflicting segments. Another 30 stimuli are V-seg only, truncated items. The full set of 120 items are presented to 34 listeners in a random order. For all tests, subjects listened at their own pace and could use a "replay" facility.

Task:

Subjects were asked for a four choice identification task: /p/, /t/, /k/ or "neither". Here, "neither" choice standed for both "no consonant" and "unidentifiable consonant" percepts.

Results :

- All original items are well recognised (99.12%). Only /py/ was occasionally identified as /ky/.
- Listeners, when asked, were unable to separate the padded stimuli from the originals while they identify easily the truncated (V-seg only) stimuli.
- 56.7% of the 1020 presentations of truncated seg-

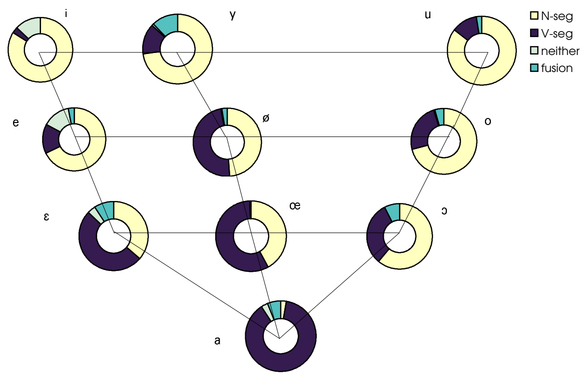


Figure 1: Results of experiment I for acoustic monsters. For each of ten French vowels, the percentage of four types of listeners responses are presented. The responses directed by the N-seg are figured in light while those directed by V-seg are dark . “neither” responses are figured in light gray and fusion (McGurk-like effect) responses in dark gray.

ments were classified as “neither”. But, when listeners do perceive a plosive, they perceived the correct one at a high rate of 84.5%. Among these correctly identified items 47% were /p/, 25% /t/ and 28% /k/. When incorrectly identified the answer was always /p/.

• Among the 2040 answers for the monsters 91.2% of the answers were distributed between consonant suggested by the N-seg or V-seg segment. The proportion is strongly depending on the vocalic substrate (see fig. 1): for the /a/ context, the elicited answer is almost always suggested by V-seg. On the contrary, this is suggested by N-seg in case of constricted vowels (/i/,/y/,/u/).

• N-seg generally tend to mask V-seg information (see fig. 2). We find that consonants rank similarly (first /k/, then /t/ and finally /p/) concerning both their N-seg or V-seg perceptual salience. For instance, most of /k/ N-seg force a /k/ percept (column 1, line 1) and /k/ V-seg resist best to perceptual overwriting by N-segs (column 2, line 4).

• Few cases of perceptive illusions of the third stop (5.2% of responses) occur in this experiment. Considering a “winnner takes all” (WTA) decision across all listeners for each stimuli, only /kɛ/ N-seg before /pɛ/ V-seg and /py/ N-seg before /ty/ V-seg were perceived as the third consonant.

• Only 3.6% of monsters presentations were labelled as “neither”. These rejected items are mostly built with N-seg segments from /p/ in front vocalic context. A posteriori verification for some listeners show that they are perceived as “no consonant”.

3.2. Experiment II

This experiment was run to study the influence of the relative amplitude of N-seg versus V-seg.

Stimuli:

Stimuli	Responses suggested by		third consonant responses	neither responses
	N-seg	V-seg		
N-seg = k	79.12%	15.88%	4.71%	0.29%
N-seg = p	34.26%	50.74%	5.15%	9.85%
N-seg = t	55.00%	38.53%	5.74%	0.74%
V-seg = k	46.18%	41.47%	6.62%	5.74%
V-seg = p	68.38%	26.32%	4.56%	0.74%
V-seg = t	53.82%	37.35%	4.41%	4.41%
overall	56.13%	35.05%	5.20%	3.63%

Figure 2: Results of experiment I for acoustic monsters grouped according to their N-seg (line 1 to 3) and according to their V-seg (line 4 to 6).

We take the 20 conflicting stimuli padded from /t/ and /k/ segments. Among them, 15 were perceived accordingly to their N-seg part so we diminish the amplitude of the N-seg in 3 steps (-6, -12 and -18 dB) in an attempt to reveal the V-seg suggested consonant. The remaining 5 stimuli were perceived as suggested by their V-seg part. Consequently, 3 steps amplification (+6, +12 and +18 dB) was performed on theirs N-segs.

16 original control stimuli were added to these 60 manipulated stimuli before randomisation in order not to bias the native perceptive space of the listeners. Only 15 listeners from experiment 1 participated in experiment 2 and 3.

Task:

In experiment 2, the subject faced the same identification task as in experiment 1. In addition, listeners were asked to choose a “sure”/”not sure” option on each phonetic decision they made.

Results :

This experiment shows that stimuli at most lose their naturalness when manipulating the amplitude of the N-seg made but the decisions of our subjects do not change significantly.

• Most responses (76.8%) are not affected by this N-seg manipulation. 11.1% of responses changed only between 0dB condition and 6dB condition (see lines 1 to 3 in fig. 3).

• These 11.1 percents yielded 6 different WTA decisions from those of experiment 1 among the 20 conflicting stimuli investigated here.

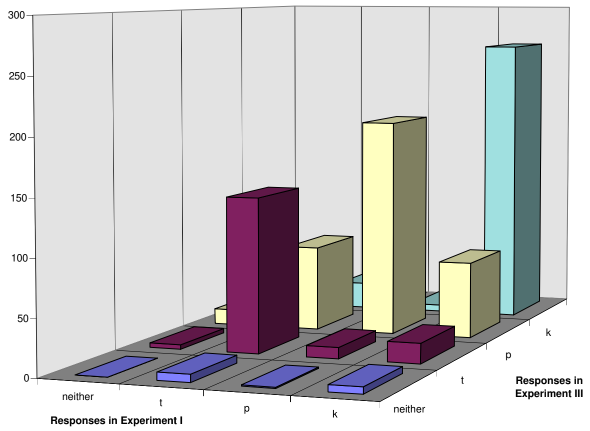
• *Categorical perception effect:* for the 3 items involving /t/ N-seg with /k/ V-seg in back vocalic context decisions are transfered from /t/ to /p/, especially when gradually diminishing the N-seg level.

• *Cues competition effect:* for 3 items, the balance between N-seg and V-seg perceptual relevance was actually inverted manipulating N-seg amplitude. They involve only intermediate aperture vowels: /ɛ/,/e/,/ø/.

• The 7.5% of responses changing between ±6dB and ±18dB conditions plus the 4.6% changing erratically never yielded different overall decisions from those at ±6dB condition. In fact, we found that the patterns of confusions in ±12dB and ±18dB conditions were not sig-

Responses in Exp. 2			Responses in Exp.1 (0dB)			
+/-6dB	+/-12dB	+/-18dB	neither	k (172)	p (23)	t (84)
k	k	k		155		14
p	p	p	1		21	14
t	t	t		2		39
k	k	neither		2		
k	t	t		1		
k	k	t		6		
t	t	neither				2
t	t	k				4
t	p	p				4
t	k	k				2
neither	neither	neither				1
neither	k	neither		1		
neither	k	k		1		
k	p	k				1
k	t	k		2		1
t	k	t		1		1
t	t	p			1	
t	p	p		1	1	
k	k	t				1

Figure 3: How responses in experiment 1 evolved when modifying N-seg energy. First 3 lines represent the mains effects. Lines 4 to 10 represent responses that changed somewhere between ±6dB and ±18dB conditions (7.5%). Lines 11 to 19 regroup responses that changed inconsistently or in an unexpected manner (4.6%).



		Responses in Experiment III			
		/t/	/p/	/k/	neither
Responses	/t/	57.2%	31.3%	09.8%	01.7%
in	/p/	04.7%	92.0%	02.8%	00.5%
Experiment I	/k/	05.0%	19.2%	74.0%	01.8%

Figure 4: Number of responses for conflicting stimuli with modified VOT against responses for the same stimuli before VOT manipulation. Most of responses don't change except that, after modification, /p/ response act as an attractor. (Incidentally, there are more /k/ responses in both tests.)

nificantly different from ±6dB condition.

3.3. Experiment III

Relative importance of N-seg versus V-seg observed in experiment 1 follow roughly the N-seg duration pattern. Longer the N-seg, greater his perceptual weight. This experiment was run to study and quantify the influence of the VOT.

Stimuli:

The duration of each N-seg of the 60 conflicting items in experiment 1 was manipulated to conform to the original V-seg consonant VOT. This duration was modified by a high-quality TD-PSOLA [3] operating on the residual of a pitch-synchronous analysis. The parameters of the LPC re-synthesis were linearly interpolated in order to best reproduce the eventual formant transitions observed in the N-seg. It was not possible to bring all N-seg from /p/ to the target VOT while maintaining the original shape of the spectral transition on a high resolution spectrogram. 5 of the 20 /p/ monsters were thus replaced with control stimuli.

Task: Same as experiment 2.

Results :

• The VOT could not entirely explain responses pattern observed in experiment 1. Among the 55 conflicting stimuli in both tests only 16 yielded different WTA percepts.

• VOT carry information for place of articulation in French to some extent. In 12 out of the 16 perceptual shifts the chosen consonant is the VOT suggested one.

• The stimuli with /p/ like VOT were almost always perceived as /p/ regardless of the spectral shape of their N-seg part. This effect produced 10 out of 16 observed shifts.

• The mean level of listeners confidence was high (mean 84.9%, std 10% in experiment 2; mean 81.6%, std 16% in experiment 3). But subjects exhibited extremely different strategies using "sure"/"not sure" option. They ranged from 100% confident to about 40% confident.

4. DISCUSSION

Previous research has demonstrated that both N-seg and V-seg of the CV continuum could be sufficient although not necessary to the perception of place of articulation of plosive sounds. Our results on French strongly support Smits et al. [9] conclusions. They show that both segments conspire to the phonetic decision in a manner heavily dependent on the following vowel: it is more like competition for /a/ and /u/ context and much like a collaboration for front vocalic context.

A short, week or even absent N-seg is an important cue for a bilabial consonant. But the relative perceptual weights of N-seg and V-seg cues in /t/ versus /k/ contrast is not sensitive to the relative energy and timing of the two segments. These results predict that models of identification such as [6, 7] based on dynamic spectral changes (intra and inter segment) will perform more efficiently then recognition models based on either V-seg only or N-seg only cues.

5. ASR EXPERIMENTS

In order to test models using integrated N-seg plus V-seg acoustic cues, we have choose to implement the algorithm described in [8].

Method :

A running spectra was calculated for each acoustic monster (60 in exp. 1, 60 in exp. 2, 55 in exp. 3) and for 276 natural CV of the same speaker (including those used for monsters construction). 20 order LPC smoothed spectrum was calculated on 14 ms windows every 5 ms for a period of 50 ms starting at the burst release (manually marked). These 11 spectra for each stimuli were Bark scaled. Then each spectrum in 200Hz-6kHz frequency range was recoded using a 7 coefficients DCT transform. Finally, each evolution of a spectral DCT coefficient over the 50 ms period was coded again using a 3 coefficients temporal DCT. Thus 7*3=21 feature vector is used to represent the running spectra of each stimuli.

The classifier is simply the Bayesian maximum likelihood one. The hypothesis of multivariate Gaussian distribution of examples is validated by good scores on the training sets (81.7% on monsters, 91.6% on natural CVs).

Results :

We train the classifier on natural CVs and test it on acoustic monsters using as reference the most-occuring response for each of them (WTA policy).

• This classifier poorly predict listeners responses : 62.9% (63.3% for experiment 1, 66.7% for experiment 2 and 58.2% for experiment 3, while chance level is about 33%).

• 78.3% of classier outputs on experiment 1 monsters deliver the N-seg consonant while only 15% point to the V-seg consonant. A similar pattern occur for experiment 2. On the contrary, classifier scores on experiment 3 monsters greatly favour the V-seg suggested consonant. Note that experiment 3 consist of a realignment of V-seg relative to burst release as in original stimuli.

• Results breakdown for the test on experiment 2 is : 80.9% /k/, 66.7% /t/, 0% /p/ correct identification. All /p/ perceived stimuli (9) were classified as /t/ as listeners actually respond in experiment 1.

• For reference, the same method perform far better on natural stimuli : 83.24%, std:5.2% on leave one speaker out cross-validation (8 speakers, 2208 CVs). Moreover, using larger DCT vector (9*5) this performance improved (88%, std:4.5%) while monster responses prediction remained the same (61.7%).

Discussion :

We conclude that this classifier, while capturing well relative invariance on natural stimuli, is not able to predict listeners responses to conflicting cues stimuli. The main reason is the high sensibility to time misalignment between training and test data. A better algorithm should be capable to search the best "temporal" matching before taking a decision on each test stimuli.

6. GENERAL CONCLUSIONS

Benchmarking automatic classifiers with respect to human responses on these conflicting stimuli should lead to robustness gains. Actual recognisers based on statistical modelling need huge training data. But this data often do not capture variability observed in real communication situations such as Lombard effects. One way to compensate for lack of sufficient statistical coverage of training data is to impose more structural constraints such as the articulatory HMM states proposed by Deng et al [4]. Another way could be to guide learning using as inputs synthetic data with mastered acoustic properties such as the ones used here. Conflicting cues stimuli are one of the extreme sets of synthetic stimuli shaping human perception performance.

REFERENCES

[1] Blumstein, S.E. and Stevens, K.N. Perceptual invariance and onset spectra for stop consonants in different vowel environment. *Journal of the Acoustical Society of America*, 67:648–662, 1980.

[2] Bonneau, A., Djezzar, L., and Laprie, Y. Perception of place of articulation of French stop bursts. *Journal of Acoustical Society of America*, 100(1):555–564, 1996.

[3] Charpentier, F. and Moulines, E. Pitch-synchronous waveform processing techniques for text-to-speech using diphones. *Speech Communication*, 9(5-6):453–467, 1990.

[4] Deng, L. and Sun, D. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. *Journal of the Acoustical Society of America*, 95:2702–2719, 1994.

[5] Dorman, M., Studdert-Kennedy, M., and Raphael, L.J. Stop-consonant recognition: release bursts and formant transitions as functionally equivalent, context-depending cues. *Perception and Psychophysics*, 22:109–122, 1977.

[6] Kewley-Port, D., Pisoni, D.B., and Studdert-Kennedy, M. Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, 73(5):1779–1793, 1983.

[7] Lahiri, A., Gewirth, L., and Blumstein, S.E. A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. *JASA*, 76:391–404, 1984.

[8] Nossair, Z.B. and Zahorian, S.A. Dynamic spectral shape features as acoustic correlates for initial stop consonants. *Journal of the Acoustical Society of America*, 89:2978–2991, 1991.

[9] Smits, R., ten Bosh, L., and Collier, R. Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. perception experiment. *Journal of Acoustical Society of America*, 100(6):3852–3864, 1996.

[10] Sussman, H.M., McCaffrey, H.A., and Matthews, S.A. An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, 90(3):1309–1325, 1991.

[11] Walley, A.C. and Carrell, T.D. Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants. *Journal of the Acoustical Society America*, 73(3):1011–1022, 1983.