COOPERATION AND COMPETITION OF BURST AND FORMANT TRANSITIONS FOR THE PERCEPTION AND THE IDENTIFICATION OF FRENCH STOPS

Adrian NEAGU, Gérard BAILLY

Institut de la Communication Parlée
46, avenue Félix Viallet 38031 GRENOBLE FRANCE
e-mail : neagu@icp.inpg.fr

ABSTRACT

In this paper, we study the influence of the vocalic context on the perception and automatic recognition of stops. In a previous perception experiment [1] using conflicting cues stimuli, we have shown that place of articulation cued by formant transitions may be overwritten by the place cued by the burst. This effect is inversely proportional to the vowel aperture. Here we give special attention to /i/ context where nor burst, nor formant transitions seem to carry rich information on place of articulation.

We present here automatic recognition experiments that confirm perception results. Taking into account both segments increase identification rates, early fusion of segmental cues performs best and most errors come from the front unrounded vocalic context.

We introduce the "burst characteristic frequency" (BF) that palliates for the poor discriminative power of the traditional cues in the front context. Moreover we present perception results showing the perceptual relevance of BF.

1. INTRODUCTION

Noise burst, hereafter called **N-seg**, and the voiced transition, hereafter called **V-seg**, both cue place of articulation while not being absolutely necessary for place perception. Recently, Smits [3] conducted a perception test to asses N-seg and V-seg perceptive weights. He proposes a model [4] that explains perception results: V-seg is more discriminant than N-seg in an open vowel context and vice-versa in a closed context. The convergence of both perception and identification experiments seems to suggest that perception focus on the most contrastive segment.

Our perceptual results on French [1] confirm this trend. We give here automatic recognition experiments aiming at the prediction of such a behavior. We show [2] that most fusion architectures give good results: they focus as expected on the most contrastive cues. However, a competitive focus can not entirely explain perceptive weights: the predictions of listeners responses for the front unrounded context are not satisfactory, as in [3, p. 3877].

In the following we bring evidence that another subtle fusion mechanism - cooperation - takes place at least in /i e/ context: the perception and the spectral analysis of the N-seg should be guided by V-seg. We will finally show that combining the competition and cooperation models both explain perceptual results and improves recognition scores.

2. PERCEPTION RESULTS

The stimuli were constructed from 30 CV stressed syllables (/p t k/ before 10 French vowels, symmetric context) pronounced by one trained male speaker. For each CV, the N-seg and V-seg were manually extracted.

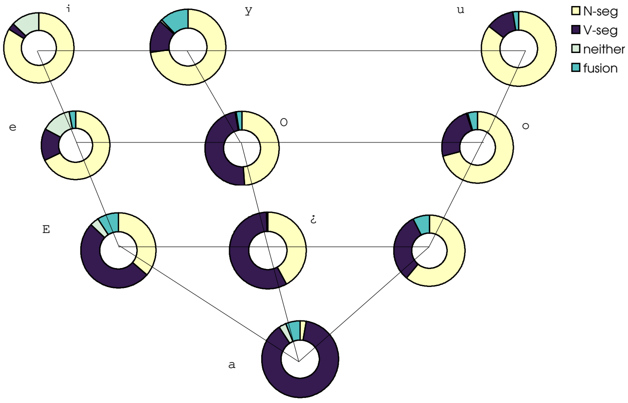

Figure 1: Variation of the N-seg versus V-seg perceptive weight according to the vocalic context. N-seg corresponding responses in light and V-seg ones in dark. McGurk-like fusion responses in dark gray and "neither" responses in light gray.

2.1. Global trends

We remind here the most impressive effect : the ratio of responses imposed by the N-seg versus responses imposed by the V-seg is heavily dependent on the vocalic context (see Fig. 1). These results are confirmed for other languages [3]. Statistics restricted to /t/ and /k/ stops exhibit even greater variation : subjects respond 100% like the N-seg in /y/ context and 100% like the V-seg in /a/ context. More details about the response pattern as well as results from perception test involving N-seg amplitude and length manipulations can be found in [1, 2].

2.2. The /ti/ versus /ki/ paradox

The perceptual dominance of V-seg in /a/ context and of N-seg in /u/ context is confirmed by tests using isolated segments [5].

By contrast, the front unrounded context stands apart. In /i/ context, like in /u/ context, the N-seg seems to mask V-seg information. The competition model would predict a salient N-seg. However, the isolated N-seg from /i/ context performs poorly for place of articulation perception, unlike N-seg from /u/ context (see Fig. 2): it is not as salient as expected for a dominant segment.

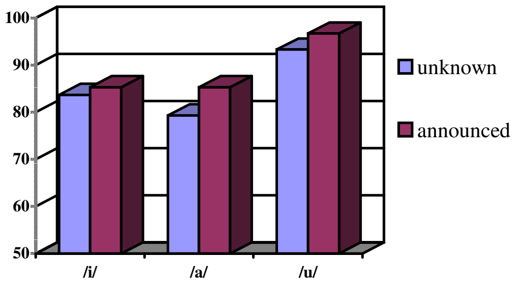

Figure 2: Stop identification from isolated N-seg, from [5] for 2 test conditions: unknown vowel (light) and pre-announced vocalic context (dark).

The way out of this paradox is suggested by gating tests results reported in [6]. Adding one or more voiced periods after the N-seg greatly reduced place of articulation confusions for /i/ and /e/ vocalic context. One should conclude that V-seg in /i/ context, while not informative by itself, is needed by human perception to read the information contained in the N-seg .

2.3. Competition and cooperation in perception

We identified 3 typical distributions for the place of articulation information in CV transitions:

- In /a/ context, V-seg is discriminant and N-seg is not.
- In /u/ context, N-seg is far more discriminant then V-seg
- In /i/ context, N-seg has the information but it need the V-seg to show it off.

We can model /a/ and /u/ associated behavior as a **competition** between N-seg and V-seg cues: perception focus on the most discriminant part. In /i/ context, human perception needs both segments. This behavior should be modeled as a **cooperation** between segments cues.

3. AUTOMATIC IDENTIFICATION TESTS

The speech corpus is an extension of the one used in perception tests. We added another 7 native French male speakers, asymmetric vocalic context and two conditions: stressed and unstressed. These add up to 2208 CV syllables.

3.1. Fusion architectures

Since both segments can cue place of articulation for stops, the automatic recognizer must deal with the fusion of data from these two sources. We tested the 3 architectures (see Fig. 3). A more detailed taxonomy of fusion models can be found in [2].

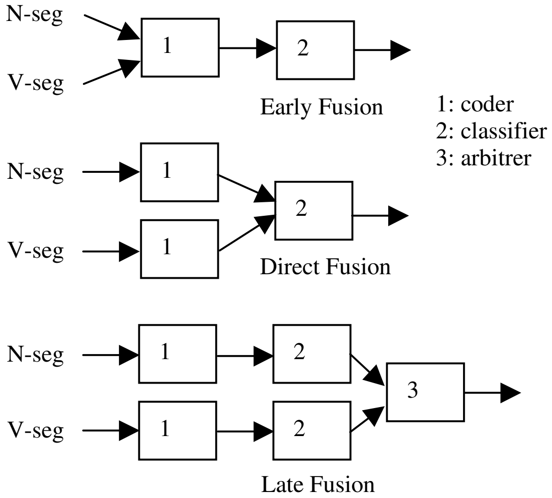

Figure 3: The 3 models we tested to fusion data from N-seg and V-seg. The coders are straight-forward feature extraction algorithms. The classifier is the only trainable part of the architecture. We always use a bayesien classifier. Among different implementations for the arbiter, we retained a simply probability multiplier.

The early fusion architecture (EF) uses a parameter set computed on a fixed length segment ignoring the underlying nature (voiced or not). The fusion takes place in the feature extraction process. Blumstein's integrated spectra is an example.

The direct fusion architecture (DF) uses parameters extracted separately from N-seg and V-seg. Here the fusion take place in the training process. Smits et al. [3] use this architecture.

The late fusion architecture (LF) uses 2 classifiers and then arbitrates between them. This is symbolic-level fusion. To our knowledge, this architecture has not been tested for stop identification.

3.2. Segments coding

We searched the best coding for both segments independently. In these pilot experiments, each parameter set is passed through a canonical discriminant analysis (CDA). For each segment, the set yielding the best stops identification score was retained.

N-seg segment is best coded by the energy of the release spectrum computed in 10 frequency bands equally spaced on a Bark scale. The spectrum is computed using a 10 ms asymmetric hamming windows starting at release (slightly longer then the minimum VOT) and is smoothed by a 20 order LPC. Different frequency scales, number of bands and LPC orders were tested. Note that our release burst spectrum, being unvoiced, differ from that computed by Blumstein [7].

V-seg segment is best coded by the F1 - F4 formant tracks, which correspond to classic coding scheme for this segment. An automatic tracking method is used. A 3 coefficient DCT model is used to smooth the amplitude and the frequency of the last 50ms of each formant track. Different track models and a global spectrogram coding of the same voiced period were tested.

Coding for the early fusion architectures needs a different approach. We chose to code a fixed length segment, longer then the average VOT, starting at the burst release. It is called the **50ms-seg** hereafter. It is best represented by a double discrete cosines transform (DCT) coding of the LPC smoothed spectrogram. This is essentially the coding method proposed by [8]. We tested also versions of Lahiri metric [9] and integrated average spectra as proposed by Blumstein [7].

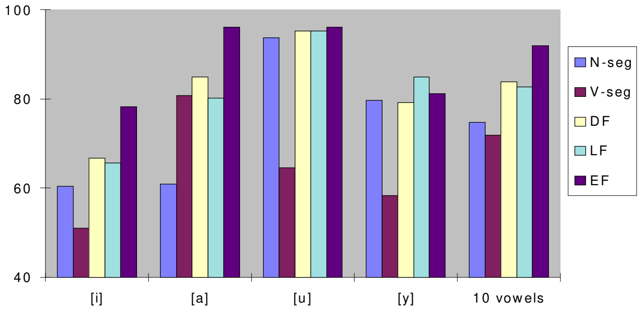

Figure 4: Vocalic context influence on the rates of place of articulation identification: 5 architectures compared. Results presented for 4 extreme vowels and averaged across 10 French vowels. Training corpus has equilibrated distributions of contexts.

3.3. Errors analysis

Previous studies did not detailed the main sources of errors. We paid special attention to the vowel effect on errors and to architecture–related limitations. Figure 4 presents identification rates for 4 vocalic contexts and averaged across the 10 French vowels. 2 non-fusion and 3 fusion architectures are compared. These results show that:

- Automatic tests confirmed perception patterns: V-seg performs better in open context while N-seg dominate the other contexts. Note that N-seg salience for /i/ is as low as for /a/
- The errors came mostly from the unrouded front vocalic context, even for the best performers. Again, we find that /i/ context is special.
- Fusion increases identification level. The early fusion is the best. Note that EF achieve its superiority in contexts where N-seg is not directly discriminantant.

4. BURST CHARACTERISTIC FREQUENCY

4.1. Mutual information

In order to understand why French /ti/ versus French /ki/ is such a frequent error, we computed normalized mutual information (NMI) patterns between release burst spectra and the place of articulation (see Fig. 5). Note the difference between vocalic contexts:

- In back and front rounded context, there are 2 information-rich frequency ranges, a /t/ zone (where /t/ spectra has statistically more energy then the /k/ spectra) and a /k/ zone.
- In front unrounded context, only one range remain. Moreover, it is a /k/ zone placed at typical /t/ zone frequency.

Due to the great contrast in the /t/ zones, a N-seg based classifier ignoring vocalic context will learn that /t/ spectra is greater somewhere. So it will tend to identify /ki/ as /ti/.

Figure 5: Normalized mutual information (NMI) between release spectrum and place of articulation for /t/ vs. /k/ contrast. NMI peaks show the most informative band in the spectrum. a) /u o/ context, b) /y ø/ context and c) /i e/ context.

4.2. Automatic identification tests

One possible solution to avoid this confusion is to force classifier to ignore /t/ zones and concentrates on /k/ zones. We call "burst characteristic frequency" (BF) the NMI peak which is the center of these variable frequency /k/ zones. BF can be computed from the vocalic context (see Fig. 6). We code the release spectrum with the energy of three 2-Bark bands around BF. This 3 coefficients set drastically reduces /t/ versus /k/ confusions made by the 10 coefficients set coding the hole spectrum (see Tab. 1). Meanwhile, it is not globally performant due to /p/ versus /t/ errors. Still, combining the 10 bands set and the BF-centered set yielded a recognition rate on a par with EF with far smaller parameter count (see Tab. 2).

Figure 6 : BF frequency plotted against first four formants position for 10 oral French vowels.

Training and testing condition	3-bands	10-bands
same context	90.3%	92.7%
different context	80.3%	39.2%

Table 1: /t/ vs. /k/ identification rates. 384 syllables. Note the under hazard level (50%) rate for cross-context validation: the release spectrum is heavily context dependent. The BF-centered band is much less variable.

Coding		learning	recognition
N-seg	3 bands	72.4%	71.8%
N-seg	10 bands	75.0%	72.7%
EF	21 coefficients	85.9%	83.2%
EF	45 coefficients	97.8%	87.9%
N-seg	10 +3 bands	87.8%	86.5%

Table 2: /p t k/ recognition rates (leave one speaker out cross-validation) and learning rates. 2208 syllables. Our early fusion (EF) is similar to Nossair [8] coding.

4.3. Perception tests

If BF centered band is perceptually relevant, diminishing its energy in a /k/ syllable should result in a /t/ perceived one. When we filtered out a 3.4 kHz band around BF in a /ki/ N-seg (see Fig. 7), 10 out of 15 listeners heard /ti/. Shortening the N-seg increase this rate to 14/15. Plauché [10] reported similar results on Spanish /ki/: all their 15 listeners perceived a /ti/.

5. DISCUSSION

In this study, we confirm the N-seg versus V-seg perceptual weight pattern for a rich vocalic system (French). We found that focus on the most discriminant segment is the principal explanation of the observed pattern. The most effective cue takes over the least effective one: it is a competition model.

Meanwhile, we found that V-seg cues in /i/ context, which are not directly discriminating for place of articulation, are, however, important in stop perception. In this context, a principal cue need a less effective one to become more effective: it is a cooperation model.

We claim that the competition model can not entirely account for the observed pattern. Actively adjusting the cues set per vowel context is necessary to explain stop confusions in /i/ context. We state that a vocal tract aware, **vowel dependent** perception mechanism coexists with the **vowel independent**, contrast based one.

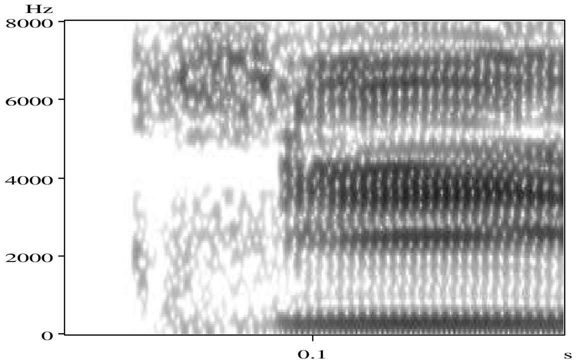

Figure 7: /ki/ syllable with stop-band filtered N-seg. This CV is perceived as /ti/.

6. REFERENCES

1. Neagu, A., Bailly, G., "Relative contributions of noise burst and vocalic transitions to the perceptual identification of stop consonants", *Eurospeech*, *Greece*, 1997, p. 2175-2179

2. Neagu, A., "Représentations phonétiques et identification des syllabes occlusive - voyelle en français", Doctoral dissertation, I. N. P. Grenoble, France, 1998

3. Smits, R., Ten Bosch, L., Collier, R., "Evaluation of various sets of acoustical cues for the perception of prevocalic stop consonants", Part I and II, *J. A. S. A., Vol.100(6),* 1996, p 3852-3881.

4. Smits, R., "Context-dependent relevance of burst and transitions for perceived place in stops: It's in production, not perception", *ICSLP, Philadelphia,* 1996, p 2470-2473

5. Bonneau, A., Djezzar, L., Laprie, Y., "Perception place of articulation of French stop bursts", *J. A. S. A., Vol.100(1),* 1996, p 555-564.

6. Kewley-Port, D., Pisoni, D. B., Studdert-Kennedy, M., "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants", *J. A. S. A., Vol.73(5),* 1983, p 1779-1793.

7. Blumstein, S. E., Stevens, K. N., "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants", *J. A. S. A., Vol.66(4),* 1979, p 1001-1017.

8. Nossair, Z. B., Zahorian, S. A., "Dynamic spectral shape features as acoustic correlates for initial stop consonants",

9. Lahiri, A., Gewirth, L., Blumstein, S. E., "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study", *J. A. S. A., Vol.73(1),* 1984, p 391-404.

10. Plauche, M., Delogu, C., Ohala, J. J., "Asymmetries in consonant confusion", *Eurospeech*, *Greece*, 1997, p. 2187-2190