Perception of time-compressed rapid acoustic cues in French CV syllables

Caroline Jacquier and Fanny Meunier

Laboratoire Dynamique Du Langage (CNRS – Univ. Lyon 2)
Institut des Sciences de l'Homme, 14 avenue Berthelot, 69363 Lyon cedex 07, FRANCE
jacquier@isc.cnrs.fr
fanny.meunier@univ-lyon2.fr

Abstract

The cognitive use of the phonetic and acoustic features still needs to be specified for speech comprehension. Several studies have established that children with language impairments like dyslexia exhibit deficits in perceiving rapid speech sounds. Our study explored the temporal encoding of acoustic cues during natural speech perception. We focused on two short attributes of speech: Voice Onset Time (VOT) and formant transitions. Bisyllabic CVCVs were constructed using stop consonants (/b/, /d/, /p/ and /t/) and vowels (/a/ and /i/). Normal hearing subjects had to identify the stimuli time-compressed acoustic cues simultaneously (Experiment 1) and separately on each cue (Experiments 2 and 3). Our results showed a non additivity of the acoustic cues and demonstrated that the VOT is a greater temporal cue than the formant transition. In addition, the redundancy of those ones is used to restore the degraded speech signal. However, both acoustic cues are needed to realize a fine perception of speech.

1. Introduction

Speech is a key tool in daily communication, and is delivered in various situations and contexts; we hear speech as a whisper, surrounded by extraneous noise or through the telephone in all of which may alter the integrity of the signal. In many situations, the speech signal is degraded by environmental sounds or sometimes even by the speaker himself. Overall and within limits speech remains understandable in spite of degradations: the cognitive system appears to be able to compensate and to reconstruct even a corrupted speech signal. However, it appears that we are not all on an equal footing faced with the perception and comprehension of degraded speech. The necessary cognitive restoration seems to rely on individual abilities [1].

Previous studies have underlined the importance of acoustic cues in speech perception [2-4]. Indeed, the speech signal is a complex combination of a wide range of acoustic cues, which have varying degrees of effectiveness for speech restoration. Therefore, measures of fine-grained speech-sound perception are potentially useful for the identification of most significant acoustic cues.

Authors have established that children with learning or language problems exhibit deficiencies in perceiving speech sounds acoustically [5]. What this means is that some rapid, brief speech sounds are poorly perceived by subjects with language disorders. However, the nature of the acoustic cues that induce cognitive deficits has still to be fully defined. In our study we explored how normal subjects deal with the degradation of rapid, brief speech sounds?

1.1. Speech Restoration

Several studies have demonstrated that language remains understandable in spite of certain acoustic degradations: man possesses the cognitive ability to restore even a speech signal which has been altered by various factors, evidenced by the fact that if a phoneme within a word is replaced by noise, subjects still perceive the word in its entirety [6]. However, correct identification is highly dependant on context and this reconstruction ability further depends on both the type and degree of the distortion. In the present study, we focused on temporal modulation in order to evaluate the importance of the acoustic cues (VOT and formant transition) in the cognitive restoration of degraded speech.

1.2. Voice Onset Time (VOT)

VOT is one of the primary acoustic cues contrasting syllable-initial stop consonants across languages both in production and perception. According to Lisker and Abramson [7], the VOT is defined as the time interval between the plosive release and the onset of voicing. If we assign a value of zero to the instant of stop release, unvoiced stops are then measured as a positive VOT, because there is a delay between the stop release and the voicing onset and, voiced stops, with a laryngeal vibration that continues from closure up to the moment of release are measured as a negative VOT. Thus, VOT values correspond to the degree of voicing. Recent studies have focused on the limits of voicing duration using experiments with a VOT continuum in order to determine the VOT value from which subjects perceived confused consonants [8].

1.3. Formant Transition

Formant transition corresponds to the rapid changes in formant frequency which occurs at the moment of release of the stop constriction. The directions of the second and third formant transitions depend both on the place of articulation and on the following vowel. The direction of formant transitions that result in a [d] percept differ between the [i] context (rising) and the [a] and [u] contexts (falling). Rapid changes in formant frequency are crucial in identifying sound segments. The transition of the second formant functions as a cue for determining the place of articulation of the plosive consonants. Serniclaes et al. [9] used a [ba]-[da] continuum in which they modified the onset of the initial frequency transitions (second and third formants). They found that children with dyslexia had an increased perception of within-category differences.

In this study, we wanted to explore the effect of rapid time-compressed acoustic cues on speech intelligibility by normal hearing subjects. We ran three experiments in which we compressed either two cues, VOT and second Formant Transition (Experiment 1) or each cue separately (Experiment 2 for VOT and Experiment 3 for Formant transition).

2. Material and Method

2.1. Experiment 1

2.1.1. Subjects

Thirty-two students, aged 18 to 32 years, participated in the experiment. They were all native French speakers with normal hearing and no language disorders. They were paid for their participation.

2.1.2. Stimuli

The stimuli used were 64 bisyllabic CVCVs and 16 VCV fillers. Four stop consonants (C) and two vowels (V) were combined to form each stimulus. For this experiment, we chose two voiced stop consonants, /b/ and /d/, and two unvoiced stop consonants, /p/ and /t/. The vowel used, V, was either /a/ or a /i/. Each consonant occurred with every other consonant in both syllable positions and with two different vowels ($4C_1 \times 4C_2 \times 2V_1 \times 2V_2 = 64$ CVCV).
The stimuli were recorded by a native male French speaker in a sound-proof cabin with a Sony ECM-MS907 microphone and saved as Windows PCM files (22 kHz, stereo, 16 bits).
The duration of the VOT and the duration of the F2 transition were segmented for each item (Figure 1). The transition was segmented from the moment of rapid change of the direction of the F2 to the steady-state part of the vowel. And VOT origin is the onset of burst thus VOT is positive or negative. The duration of both acoustic cues was time-compressed according to four experimental conditions:

- 100% = original duration
- 50% = 50% of original duration
- 25% = 25% of original duration
- 0% = totally cut

We time-compressed the acoustic cues using the Pitch-Synchronous Overlap Add (PSOLA) time scaling technique, used in the Praat software. With Praat, segmented parts of the waveform can be time-compressed, while the remainder of the waveform remains unaffected. In this way, each syllable can be selectively time-compressed.

2.1.3. Procedure

The participants were seated in a silent room facing a computer monitor. The stimuli were delivered binaurally via headphones (Beyerdynamic DT 48, 200Ω), and they were randomized across subjects. The subjects were informed that a speech signal, though not necessarily a word, was to be emitted.
The participants then had to type on a computer keyboard whatever they heard. They were trained with a control list before the start of the actual experiment.

2.2. Experiments 2 and 3

2.2.1. Subjects

Two groups of sixteen students, aged 18 to 23 years (Experiment 2) and aged 19 to 23 years (Experiment 3) participated in these experiments. They were all native French speakers with normal hearing and no language disorders. They were paid for their participation. None of them took part in Experiment 1.

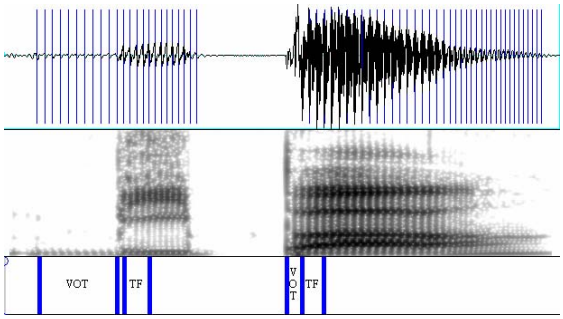

Figure 1: [bipa] oscillogram and spectrogram.
VOT of [b] is negative and VOT of [p] is positive.
(VOT=Voice Onset Time; TF=Formant Transition)

2.2.2. Stimuli

The same 64 bisyllabic CVCVs used in the first experiment were used in Experiments 2 and 3. However, in Experiment 2, only the duration of VOT was time-compressed whereas in Experiment 3, it was only the duration of formant transition that was time-compressed.

2.2.3. Procedure

The procedure for Experiments 2 and 3 was identical to that of the first experiment.

3. Results

3.1. Experiment 1

We computed the identification rates for stimuli across subjects, for both consonants and vowels. The mean identification rate for stimuli was 55.4%. Overall, vowels were better conserved and identified than consonants. Moreover, according to their position (first or second syllable), the intelligibility of the consonants was differentially modulated by time-compression and, we observed that the first consonant was less identifiable than the second (Table 1).
A two-way ANOVA including Position (first syllable S1, second syllable S2) and Conditions of time-compression (4) factors showed a significant main effect for both the position ($F(1,31)=28.97$, $p<.05$) and time-compression conditions ($F(3,93)=652.32$, $p<.05$) and an interaction effect ($F(3,93)=5.75$, $p<.05$). Identification of the consonant seemed therefore to depend on its position in the item. All time-compression conditions were also seen to have an effect on the identification rate. We also observed that the S1 identification rate was significantly different at 100% and 50% ($p<.05$) whereas this was not the case for S2. Thus, time-compression made it more difficult to restore attack consonants than intervocalic consonants, and the greater the compression, the greater the difficulty.
A three-way ANOVA analysis including Voicing (voiced or unvoiced), Place of articulation (labial or dental) and Condition (4) factors was also performed. For S1, we observed significant main effects for Voicing ($F(1,31)=544.50$, $p<.05$), for Place of articulation ($F(1,31)=9.62$, $p<.05$) and for Condition ($F(3,93)=409.19$, $p<.05$). A noteworthy result was that voiced stop consonants were better identified than unvoiced stops at 50%, 25% and 0%. Labial consonants were better identified than dental at 25% and 0%. Analysis of the

subjects' errors showed that unvoiced stops were not confused with another consonant but were in fact totally lost. Similarly, for S2, we observed significant main effects for Voicing ($F(1,31)=22.19$, $p<.05$), for Place of articulation ($F(1,31)=8.03$, $p<.05$) and for Condition ($F(3,93)=356.71$, $p<.05$). In opposition to S1, here the unvoiced stop consonants were better identified than voiced stops at 25% and 0% whereas labial consonants were better identified than dental at only 0%. The /p/ consonant was the best identified whereas the /b/ consonant was the worst identified. Another interesting result was the confusion of /b/ with the /l/ liquid consonant, most of the time within an /i/ context. We noted that the /d/ consonant was also recognized as an /l/.

Overall, we noticed a wide variability in performance between individuals. The variance in the performance of the thirty-two subjects was largest at 25% for both consonant positions (Figure 2): for S1, performances range from 94% for the best subject to 25% for the worst one. This variance may reflect a differing cognitive ability to restore degraded speech as all subjects reported having no known language or hearing disorders. However, differences in hearing abilities, though not pathological, could explain these results. Further exploration is needed to check this point.

	C_1	C_2	V_1	V_2	Items
Exp. 1 **VOT - / TF -**	70.4	77	98.9	99.1	55.4
Exp. 2 **VOT - / TF +**	87.3	90.1	99.8	99.8	78.9
Exp. 3 **VOT + / TF -**	93.6	97.8	100	99.9	91.2

Table 1: Mean Identification rates for Experiments 1, 2 and 3. Values are expressed as percentages, the sign (-) corresponds to time-compression and the sign (+) corresponds to original duration.

3.2. Experiment 2: VOT time-compression

Overall, the mean identification rate was higher than in Experiment 1. Vowels were still better identified than consonants but the position effect remained (Table 1).
A two-way ANOVA showed significant main effects for Position ($F(1,15)=6.04$, $p<.05$) and for Condition ($F(3,45)=125.26$, $p<.05$). The identification rate was significantly different only for the 0% condition ($p<.05$). Thus, VOT has to be totally removed to induce difficulties in cognitive restoration. Otherwise, time-compression had no effect on the identification rate.
In a three-way ANOVA, for S1, we observed significant main effects for Voicing ($F(1,15)=39.68$, $p<.05$), for Place of articulation ($F(1,15)=6.90$, $p<.05$) and for Condition ($F(3,45)=65.85$, $p<.05$). As in Experiment 1, we observed that voiced stop consonants were better identified than unvoiced stops at 50%, 25% and 0%. Labial consonants were better identified than dental ones at 25% and 0%. The subjects' errors showed that unvoiced stops were mostly totally lost. Similarly for S2, we observed significant main effects for Voicing ($F(1,15)=10.53$, $p<.05$), and Condition ($F(3,45)=58.07$, $p<.05$) but no significant main effect for Place of articulation ($F(1,15)=2.50$, n.s.). However, we noted contrasting results to S1, as here the unvoiced stops were better identified than voiced stops only at 0%. The /p/ consonant was still the best conserved, whereas the /b/ and /d/ consonants were confused with the /l/ liquid consonant at 0%.
Furthermore, the largest inter-individual variability was observed only at 0%. When the VOT is altered, this may reveal redundancy of the acoustic cues.

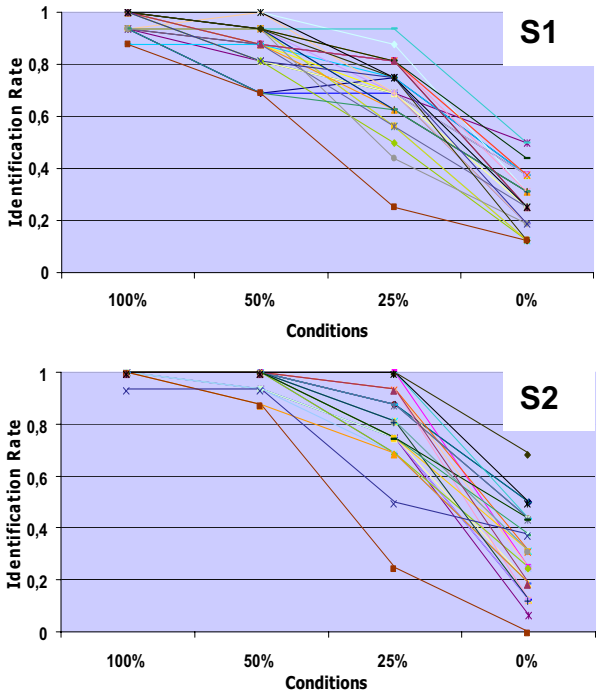

Figure 2: Inter-individual Variability for the 32 subjects in Experiment 1. Each curve gives the performance of each subject for each condition (100%, 50%, 25% and 0%). Graph S1 shows the inter-individual variance in performance for the first syllable and Graph S2 for the second syllable.

3.3. Experiment 3: Formant Transition time-compression

In Experiment 3, the mean identification rate was higher than in the former ones. A ceiling effect seemed to appear. However, vowels were still better identified than consonants. The position effect was still demonstrated (Table 1).
A two-way ANOVA showed significant main effects for both Position ($F(1,15)=36.54$, $p<.05$) and Condition ($F(3,45)=24.68$, $p<.05$). We also noted that the identification rate was significantly different only for the 0% condition ($p<.05$). It therefore appears that formant transition has to be totally removed to induce difficulties in cognitive restoration since time-compression had no effect on the identification rate.
In a three-way ANOVA, for S1, we observed significant main effects for Place of articulation ($F(1,15)=19.29$, $p<.05$) and Condition ($F(3,45)=19.40$, $p<.05$) but no significant main effect for Voicing ($F(1,15)=1.67$, n.s.). We also observed a significant main effect for S2 of Place of articulation ($F(1,15)=12.45$, $p<.05$), and Condition ($F(3,45)=11.52$, $p<.05$) but not for Voicing ($F(1,15)=3.85$, n.s.). In both positions, labial consonants were better identified than dental ones. We would point out that /d/ was most often confused with /b/ indicating that this was an error in the place of articulation.
We also noted no significant inter-individual variance maybe because of a correct response ceiling. The task seemed too easy; since degradation did not modify performance. This lack of variance may reflect redundancy of the acoustic cues that are used during speech identification.

4. Discussion and Perspectives

In this paper, we studied the effect of time-compression of rapid acoustic cues on speech intelligibility by normal hearing

subjects. We compressed either both cues (Experiment 1: Voice Onset Time and Formant Transition) or each cue separately (Experiment 2 for VOT and Experiment 3 for Formant transition). Overall, vowels were better conserved than consonants and the attack consonant was less identifiable than the intervocalic consonant. Globally, labial consonants are better reconstructed than dental ones even if specific observations depend on manipulated cues. For the first syllable, voiced stops are better identified than unvoiced ones whereas the inverse effect is observed for the second syllable. However, manipulating cues separately modulates these effects. Indeed, when only the Formant Transition is time-compressed, performances are still very high until total suppression of the signal segment. This result underlies the redundancy of features in the speech signal that are implicated in the cognitive restoration of degraded speech.

It is worth noting that even when both cues were manipulated performances was still very high up to 50% of the original duration. However, when both cues were removed, individual performances go from 0% up to around 50% (depending on the modified syllable). The largest variance in inter-individual performances was observed in the 25% condition. While some subjects are very good at reconstructing (more than 80% of reconstructed syllables), other individuals saw their performance fall drastically. In Experiments 2 and 3, the largest variance is at 0%. This latter result can be explained by the information remaining in the undamaged acoustic cue. As a result of the redundancy of temporal acoustic information, the cognitive processes of restoration can be activated and the right syllable retrieved.

Overall and in accordance with the data in the literature, vowels are better conserved than stop consonants. The steady-state parts of the vowels are changed less than the waveform of consonants. Even with time-compression of formant transition, this classic effect is preserved.

Concerning the differences in performance observed given the position of syllables, the difference in length classically described in French - intervocalic consonants (S2) being longer than attack consonants (S1) - could explain why subjects committed fewer errors for the S2 identification than for the S1 identification. Information given by the final vowel transition of the first syllable could further help in reconstruction of the second syllable.

On the other hand, the acoustic context could also play a role on the modulated voicing effect between S1 and S2. Within a voicing context (S2) unvoiced stops emerge whereas within a silent context (S1) voiced stops emerge in their turn. This may be explained by salient consonants. Another explanation is that time-compression of unvoiced stops in first syllables induces a loss of significant information from VOT duration since when unvoiced stops (/p/, /t/) are lost, they are not confused with another consonant. However our results show great confusion of voiced stops when they are in the second position, (/b/, /d/) being confused with the /l/ liquid consonant, which may explain the reduced performance in the identification of voiced stops compared to unvoiced stops. The /d/ mistakes can be easily explained by mode confusions when there were also place confusions seen for the /b/ mistakes. Early studies using English words have established that dental stops are often confused with liquid consonants (e.g., /rider/). At the articulatory level, when a dental stop is accelerated the release of constriction is reduced and it resembles the friction of liquid. An acoustic description could give details on the perception of /b/ with /l/.

Generally, the superiority of labial identification can be explained for the first syllable by the greater loss of the /t/ than /p/ (Experiments 1 and 2) and for the second syllable by a lesser, but still important loss of the /t/ and the confusion of /d/ with /l/ (Experiment 1). However, in Experiment 2, the place of articulation effect on the second syllable was not observed because of a better /t/ identification. Moreover, in Experiment 3, labial identification was also better than dental because of the great confusion of /d/ with /b/. This error on the place of articulation may reflect the distorted formant transition effect.

Overall, results allow us to identify the importance of each cue in speech restoration. An unchanged formant transition improves /d/ identification (Experiment 2) whereas the original VOT duration improves the identification of the majority of consonants (Experiment 3).

We conclude that acoustic cues are not additive features and VOT is a greater temporal cue than formant transition and that their redundancy is used to restore the degraded speech signal.

In order to correlate inter-individual variability with the subjects' cognitive performance, we plan to carry out an evaluation of reading skills and audiometric tests. The differences across subjects could be due as much to language deficits as hearing disorders. Moreover, an acoustic measurement study will be achieved to clarify the remaining obscure points.

5. Acknowledgments

The authors would like to thank Jalaleddin Al-Tamimi, Claire Delle Luche, Emmanuel Ferragne, Egidio Marsico and François Pellegrino for their help and their contribution. This research has been funded by the EMERGENCE program of the French *Région Rhône-Alpes*.

6. References

[1] Meunier, F., Cenier, T., Barkat, M., and Magrin-Chagnolleau, I., "Mesure d'intelligibilité de segments de parole à l'envers en français", *XXIVèmes Journées d'Etude sur la Parole, Nancy, 2002.*

[2] Serniclaes, W., "Etude expérimentale de la perception du trait de voisement des occlusives du Français", *Ph.D. Dissertation, Université Libre de Bruxelles, 1987.*

[3] Kent, R. D., and Moll, K. L. (1969). Vocal-tract characteristics of the stop cognates. *J. Acoust. Soc. Amer., 46(6), 1549-1555.*

[4] Lisker, L., and Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech, 10, 1-28.*

[5] Tallal, P., and Piercy, M. (1974). Developmental aphasia: rate of auditory processing and selective impairment of consonant perception. *Neuropsychologia, 12(1), 83-93.*

[6] Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167(917), 392–393.*

[7] Lisker, L., and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word 20, 384-422.*

[8] McMurray, B., Tanenhaus, M. K., and Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86, B33-B42.*

[9] Serniclaes, W., Sprenger-Charolles, L., Carré, R., and Demonet, J. F. (2001). Perceptual discrimination of speech sounds in developmental dyslexia. *J. Speech Lang. Hrng. Res., 44, 384-399.*