6th International Conference on Spoken
Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

ON AUDITORY-PHONETIC SHORT-TERM TRANSFORMATION

René Carré, Liliane Sprenger-Charolles**, Souhila Messaoud-Galusi***, Willy Serniclaes*****

*ENST-CNRS, 46 rue Barrault, 75634 Paris cedex 13
**CNRS (UMR 8606) et Université René Descartes, Campus CNRS Hôpital de Villejuif,
7 rue Guy Moquet BP 8, 94801 Villejuif
***Institut de Phonétique, 19 rue des Bernardins, 75005 Paris
****Laboratoire de Statistique Médicale, Ecole de Santé publique, CP 598, ULB,
808 route de Lennik, 1070 Brussels, Belgium

ABSTRACT

In a previous experiment, we showed that the vowel-vowel token [ai] is perceived by adult French listeners as /ai/ for transition durations between the 2 vowels lesser than 200 ms (50% of the response) and as /0 ☎8/ for larger durations. Recall that the /ai/ formant trajectory in the F1-F2 plane crosses the region of the vowel /☎/. Further, the 200 ms /0 8/-/0 ☎8/ perceptual boundary can be related to syllable duration. The same experiment was tested with children of 6.5, and 13 years old. Six tokens [ai] with different duration transitions (50, 100, 150, 200, 250, 300 ms) were randomly presented 10 times each. The question was: "do you hear 2 or 3 sounds?". Six and half years old children predominantly perceived 2 vowels and their responses were only slightly affected by transition duration. This response pattern decreased progressively with age for large duration transition tokens. At 13 years, the results were closed to the adults ones. At this point of our research, we may suppose that, at the beginning of the acquisition process, the perception is holistic, i.e. global; then, progressively with age and linguistic environment (probably also through reading acquisition), an auditory-phonetic working short term memory of syllabic duration is set up according to the syllabic structure of the French language. This memory could be used to transform the speech signal into symbolic representation (phonemes, features gestures, ..).

1. INTRODUCTION

In previous experiments, synthesized vowel-to-vowel token were automatically obtained using a deductive approach. Two criteria to the deformation of an acoustic tube of 17.5 cm length were applied: 1) maximum acoustic contrast, and 2) efficiency (or minimum effort: a small gesture deformation should lead to a large acoustic effect). In fact, these minimal constraints have formed the point of departure for the derivation of the distinctive region model (DRM) [1]. In this model, an asymmetrical behavior is observed: whereas a front constriction is automatically associated with a back cavity and vice-versa, a central constriction will be automatically associated with two lateral cavities. Acoustically efficient places of articulation also automatically follow from the model [2]: Although these model places are obtained without referring to any knowledge about articulatory observations, they nevertheless coincide with the places used to produce vowels and consonants. Figure 1 shows a simplified DRM model used to produce [ai] tokens. The transversal deformation gestures act anti-symmetrically on back and front part of the tube.

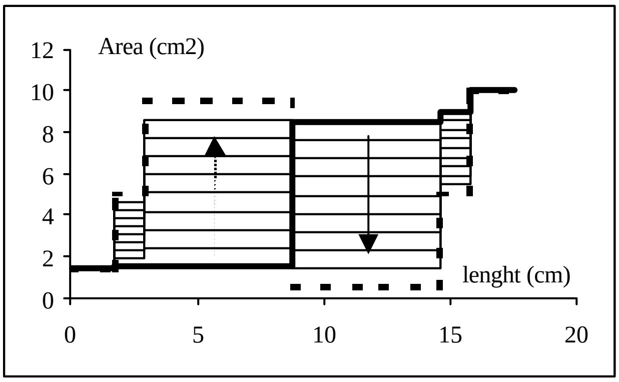

Figure 1: Crude vowel-to-vowel area function configurations in [ai] production obtained with the DRM model. The source is at the closed end of the tube (left part), the output at the open end (right part). The displacement gesture of the constriction from back to front is transversal.

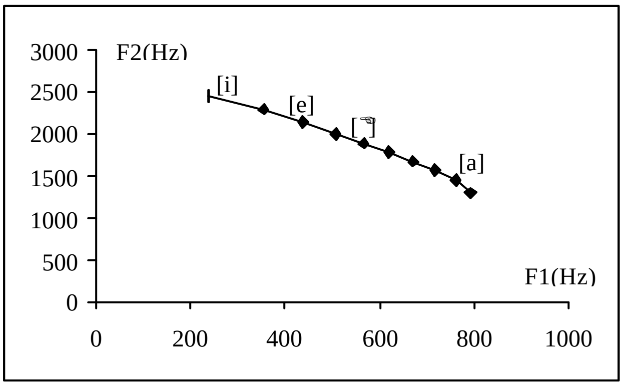

Figure 2: [ai] formant trajectory in the F_1-F_2 plane. The trajectory crosses the [☎] and the [e] region.

When the model produces an [ai] token, the formant trajectory crosses the regions of the French vowels [[illegible]] and [e] (see figure 2), although the presence of these vowels is not heard in normal speech.

In an experiment described in [3], we examined the percept generated by [ai], with gesture duration ranging from 50 to 300 ms in 50 ms steps. The duration of the first vowel was 100 ms and that of the second vowel 150 ms. Figure 3 illustrates the results of this experiment, with a single French listener as subject. The vowel complex /ai/ was perceived with gesture durations between 50 but around 200 ms, then an /a[illegible]i/ percept was reported. That is, at these longer durations an additional intermediate vowel, /[illegible]/, was heard, despite the absence of a segmental marker. Figure 4 shows the results for 4 French listeners: If the boundary is more or less at the same place, on the other hand, the standard deviation is larger. It means that the boundary depends on the subject. The test was extended to different vowels with the same result [4].

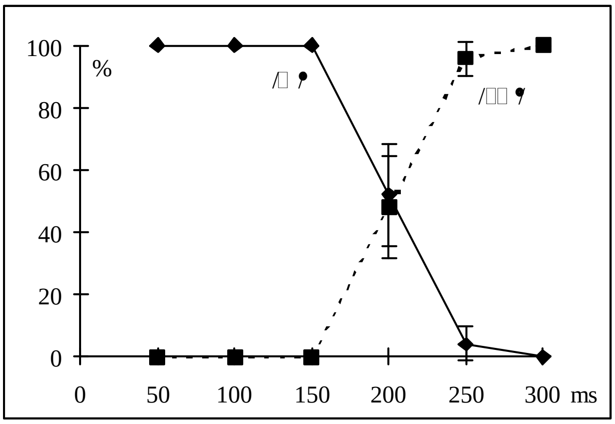

Figure 3. Percentage (mean and standard deviation values) of the percept /ai/ as a function of transition duration for one French listener. For long transition duration, an extra intermediate vowel is perceived: the percept is /ai/ for transition duration between 50 and around 200 ms. For transition duration more than 200 ms, /[illegible]/ is perceived. The standard deviation is relatively small.

The same experiment was submitted to English listeners [4] Boundaries are observed at around 400 ms. From the results of this set of experiments, it is tempting to speculate that the segmentation of V-V sound streams would be performed according to the listener's phonetic experience [5] of extracting language specific attributes of speech sounds. Evidently, the results are speaker dependant. Thus, such an experiment could inform on aspects of the process of language acquisition. So, it was decided to submit the test to children being 6.5 and 13 years old.

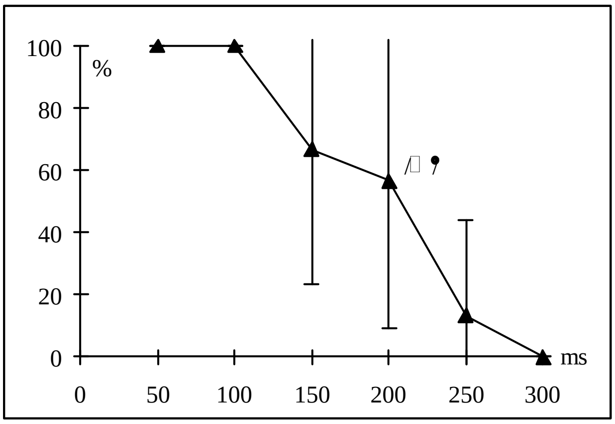

Figure 4. Percentage (mean and standard deviation values) of /ai/ as a function of transition duration for four French listeners. As expected, the standard deviation is larger compared with a single listener (see figure 3).

2. EXPERIMENT

The synthetic [ai] items were obtained with the DRM model controlled by the tongue gesture (figure 2). The time course interpolation for the transition between two adjacent vowels followed a linear function. The corresponding formant frequencies were calculated using the Badin and Fant algorithm [6] taking into account all the tract losses (heat, viscosity and radiation) and the wall vibrations. The first three formants obtained were used to control a cascade formant synthesizer. The frequencies of F_4, F_5, and F_6 were fixed at 3800, 4500 and 5500 Hz respectively. The bandwidths of the first six formants were fixed at 70, 110, 150, 200, 250 and 350 Hz respectively. The signal sample rate was 16 kHz and the formant parameters were updated every 10 ms. The voice source was represented by series of pulses having width of 0.1 ms shaped by a second-order glottal filter (F=100 Hz, BW=300 Hz). F0 varied linearly from 120 Hz at V_1 onset to 130 Hz at V_1 offset and to 100 Hz at V_2 offset. For each experiment, a set of tokens was synthesized with varying transition characteristics, i.e. 50, 100, 150, 200, 250, 300 ms. Figure 5 shows the three first formants in the time domain with a transition duration of 100 ms. A series of stimuli for each perceptual test consisted of randomized 10 repetitions of the individual tokens from the set. The stimuli were presented through earphones at a comfortable listening level. All the experiments were run on PC computer. The results for each token were expressed as mean percent identification and its standard deviation over the subjects.

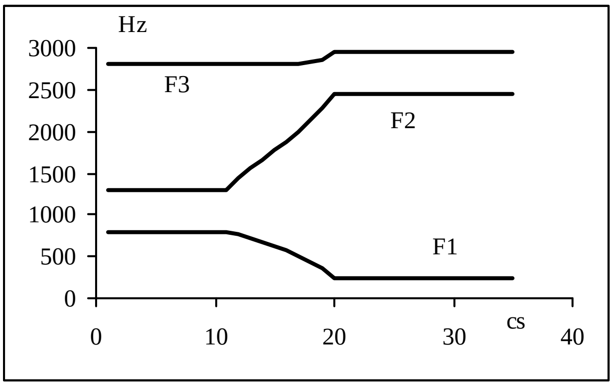

Figure 5: Formant frequency variations in the case of [ai] production. The duration of the first vowel is 100 ms, the duration of the second vowel is 150 ms and the duration of the transition is here 100 ms.

Only the children who met the following criteria were enrolled: 1. French as native language, 2. No language or motor problem, 3. Average or above average verbal and non-verbal IQs, average readers. 17 of them were 13 years old and 29 were 6.5 years old at the moment of the experience. The question was: “Do you hear 2 or 3 sounds?”.

3. RESULTS

The results of the perception are showed in figure 6. The responses obtained with the 13 year old children are closed to the adult’s ones (figure 4). But the six and half years old children predominantly perceived 2 sounds and their responses were only slightly affected by transition duration. It is also noted that the standard deviation is very high especially for the 6.5 years old children (the mean standard deviation for the six measurements are 18.3 for the 13 years old children which is comparable to the one obtained for adults-20.3, and 28.4 for the 6.5 years children). It means first that the results depend on the subject and second that, at 6.5, either it is difficult to perform the test or the process of acquisition is quite fast. This should be clarified by looking at individual results (standard deviation).

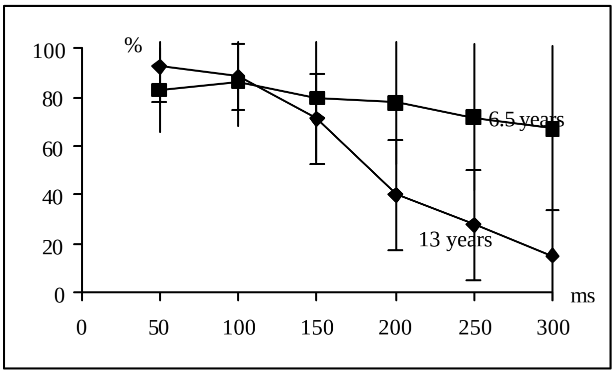

Figure 6: Percentage of perception of two sounds for children from two different ages (six and half and thirteen years).

4. CONCLUSIONS

From our set of results, it is tempting to hypothesize that the segmentation of V-V sound streams would really be performed according to the listener’s phonetic experience [5] of extracting language specific attributes of speech sounds. Exposure to the ambient language would lead to progressive acquisition of syllabic segmentation in French (about every 200 ms), and of inter-stress segmentation in English (about every 400 ms). Segmentation relies on acoustic landmarks such as a rapid spectral change or a rapid prosodic variation [7]. But in the case of our experiments, there is no specific acoustic event (landmark) corresponding to syllable or inter-stress boundary. Thus, it may be assumed then that the listener, according to his phonetic experience, integrates the acoustic information over the expected duration of the relevant segmental unit. For French and English adult subjects, our interpretation is coherent with the syllabic hypothesis initially proposed by Melher [8], and improved by Cutler et al. [9] (see also[10, 11]). They show that French speakers perform syllabic segmentation whereas English speakers do not. They show stress-based segmentation. If this hypothesis is right, an auditory-phonetic short-term transformation could be developed with age as shown by the results obtained with children and those obtained with adults. This auditory-phonetic short-term transform cannot be compared either with the auditory echoic memory (keeping the acoustic characteristics of around two seconds of the past speech signal) or the phonetic memory which is a pre-lexical memory keeping the succession of the 5 to 7 last past recognized syllables [12, 13].

In Figure 7, a schematic representation of the auditory signal to phonological unit transform is proposed. The two-second echoic memory (around seven syllables) would correspond to a pre-lexical storage. Only the right hemisphere processing (represented by the right path in figure 7) would be used at the beginning of the language acquisition. At this stage, words would be represented in terms of signal parameters in the lexicon memory. Notice that this holistic representation would occupied a relatively large amount of memory space.

With the increase of the number of words to be acquired for communication needs, the memory burden would become prohibitive. This would lead to develop common symbolic representations of sounds which share the same phonological properties (phonologisation). These representations would e located in the left hemisphere (left path in figure 7) and would progressively develop during the lexicon explosion (at around 2 years). This left path is much less costly in terms of comparisons and lexical memory size and thus needs less effort. In this frame, Then our results would suggest that the size of the short term transform can be reduced (between 6 and 13 years?) according to the syllabic or inter stress duration due to language structure. They also suggest that reading acquisition may contributes to the development of this analytic skill [14].

Such a scheme shows the importance of the development of the signal/symbol transform.

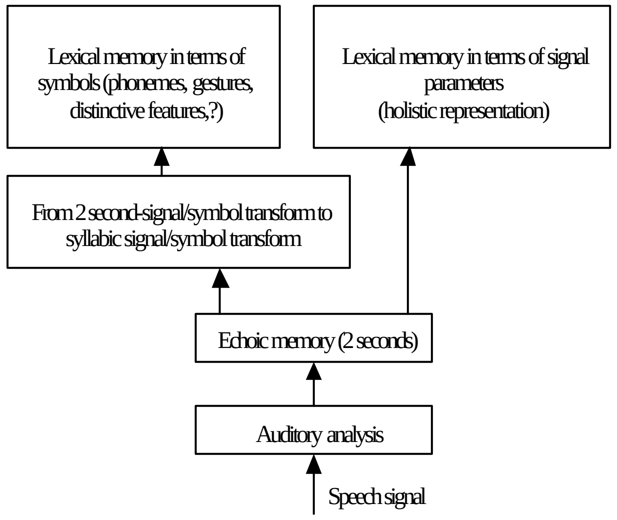
Lexical memory in terms of symbols (phonemes, gestures, distinctive features,?)
Lexical memory in terms of signal parameters (holistic representation)
From 2 second-signal/symbol transform to syllabic signal/symbol transform
Echoic memory (2 seconds)
Auditory analysis
Speech signal

Figure 7: Schematic representation of the auditory signal/symbol transform

However, such explanations deserve further studies to offer more solid interpretations. It is needed to increase the field of the study to children with different ages and also with English children.

ACKNOWLEDGMENT

This study is supported by the Cognitic Program of the French Research Ministry.

5. REFERENCES

[1] M. Mrayati, R. Carré, and B. Guérin, "Distinctive region and modes: A new theory of speech production," *Speech Communication*, vol. 7, pp. 257-286, 1988.

[2] R. Carré and M. Mody, "Prediction of Vowel and Consonant Place of Articulation," Proceeding of the Third Meeting of the ACL Special Interest Group in Computational Phonology, SIGPHON 97, Madrid, 1997.

[3] R. Carré, "Perception of coproduced speech gestures," Proc. of the 14th Int. Cong. of Phonetic Sciences, San Fransisco, 1999.

[4] R. Carré, W. A. Ainsworth, P. Jospa, S. Maeda, and V. Pasdeloup, "Perception of vowel-to-vowel transitions with different formant trajectories,",, submitted.

[5] K. Johnson, "Speaker perception without speaker normalization. An exemplar model," in *Talker Variability in Speech Processing*, K. Johnson and J. W. Mullennix, Eds. New York: Academic Press, 1997, pp. 145-165.

[6] P. Badin and G. Fant, "Notes on the vocal tract computations," *KTH, STL-QPSR*, vol. 2-3, pp. 53-107, 1984.

[7] K. Stevens, "Evidence for the role of acoustic boundaries in the perception of speech sounds," in *Phonetic Linguistics. Essay in Honor of Peter Ladefoged*, V. A. Fromkin, Ed. Orlando: Academic Press, 1985, pp. 243-255.

[8] J. Melher, J. Segui, and U. Frauenfelder, "The role of the syllable in language acquisition and perception," in *The cognitive Representation of Speech*, T. Myers, J. Laver, and J. Anderson, Eds. Amsterdam: North Holland, 1981.

[9] A. Cutler, J. Mehler, D. G. Norris, and J. Segui, "The syllable's differing role in the segmentation of French and English," *Journal of Memory and Language*, vol. 26, pp. 480-487, 1986.

[10] A. Cutler, J. Mehler, D. Norris, and J. Segui, "Limits on bilingualism," *Nature*, vol. 340, pp. 229-230, 1989.

[11] A. Cutler, J. Melher, D. G. Norris, and J. Segui, "The monolingual nature of speech segmentation by bilingual," *Cognitive Psychology*, vol. 24, pp. 381-410, 1992.

[12] D. B. Pisoni, "Auditory short-term memory and vowel perception," *Memory & Cognition*, vol. 3, pp. 7-18, 1975.

[13] J. E. Cutting and D. B. Pisoni, "An information-processing approach to speech perception," in *Speech and Language in the Laboratory, School and Clinic*, J. F. Kavanagh and W. Strange, Eds. Cambridge: The MIT Press, 1978, pp. 38-72.

[14] A. Fowler, "How early phonological development migth set the stage for phoneme awareness," in *Phonological processes in literacy. A tribute to Isabelle Y. Liberman*, S. A. Brady and D. P. Shankweiller, Eds. Hillsdale, N.J.: Lawrence Erlbaum Associated, 1991, pp. 97-117.