Binding of speech syllables when segregation occurs

Marion David[1], Mathieu Lavandier[2], Nicolas Grimault[3] and Andrew J. Oxenham[4]

[1]Universität Oldenburg, Department für Medizinische Physik und Akustik, Exzellenzcluster Hearing4All, Carl-von-Ossietzky-Strasse 11, 26129 Oldenburg, Germany

[2]Univ Lyon, ENTPE, Laboratoire Génie Civil et Batiment, Rue Maurice Audin, 69518 Vaulx-en-Velin Cedex, France

[3]Centre de Recherche en Neurosciences de Lyon, Université Lyon 1, Cognition Auditive et Psychoacoustique, Avenue Tony Garnier, 69366 Lyon Cedex 07, France

[4]Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455, USA

Abstract.

Two previous studies [David et al., 2017, Hear. Res. 344, 235-243; David et al., 2017, J. Acoust. Soc. Am. 142(3), 1674-1685] have investigated the segregation of speech syllables made of a fricative consonant and a vowel, referred to as CV tokens. The first study explored the segregation of such syllables based on fundamental frequency differences. The second study explored the segregation of the CV tokens based on localization cues, especially the spectral cues in the median plane. Both studies found that segregation can be observed based on F0 and on spectral cues. Interestingly, it was found that the whole CV token remains grouped even when segregation occurs based on cues that affect only one part of the CV: F0 differences affect mostly the vowel part, whereas coloration in the median plane is effective mostly at high frequencies, selectively affecting the consonant part. The mechanisms that allow the CV to remain grouped under such circumstances remain unclear. The present manuscript reviews the results of these two studies and provides some suggestions as to how such binding might occur.

1. Introduction

In order to understand a target voice amid a background noise (e.g., in a cocktail party; Cherry, 1953), the target must be grouped into one single auditory stream and segregated from the other sound sources. Such auditory object formation relies on our ability to organize competing sound sources into coherent streams (Bregman, 1990).

Auditory stream segregation and integration have been studied using both speech and non-speech sounds (for reviews, see Moore and Gockel, 2002; 2012). It has been shown that differences in fundamental frequency (F0) provide important segregation cues for pure and complex tones (Miller, 1957; van Noorden, 1975; Vliegen and Oxenham, 1999). In addition, sound localization cues, including interaural time and level differences (ITDs and ILDs, respectively) and monaural spectral

(or coloration) differences (Blauert, 1997; Wightman and Kistler, 1992), can induce stream segregation of pure and complex tones (Gockel et al., 1999; Sach and Bailey, 2004; Stainsby et al., 2011), as well as frozen speech-shaped noises. With some exceptions (Gaudrain and Grimault, 2008; Gaudrain et al., 2007; Hartmann and Johnson, 1991), studies of auditory streaming have tended to use sequences of single repeated sounds, making it difficult to generalize the results to real-world situations, where sounds are complex, constantly changing, and not fully predictable. The aim of the studies reviewed in this manuscript (David et al., 2017a, b) was to apply a streaming paradigm to stimuli that are closer to real-world speech, by using random sequences of consonant-vowel (CV) syllables with natural spectro-temporal variability. Such stimuli could be segregated based on differences in either F0 or simulated spatial position, despite variations along other spectro-temporal dimensions.

2. Rationale

Two sequences of sounds can be perceived as integrated – if the perceptual acoustical difference between the sequences is small enough – or as segregated – if the perceptual difference between the sequences is large enough. Auditory scene analysis is governed by both voluntary and obligatory mechanisms, depending on the task the listeners have to complete. Voluntary streaming corresponds to tasks where the listeners try to hear out a target sound from a mixture. Conversely, obligatory streaming corresponds to situations where the listeners attempt to bind the sounds into a single stream but fail to do so (Bregman, 1990).

Two interleaved sequences, A and B, were presented to the listener (cf. Fig. 1). The difference between the sequences was based either on fundamental frequency or on simulated spatial position in the horizontal or median plane. Here we discuss only the results in the median plane, as these are the ones where the spectral cues are superimposed on the inherent spectro-temporal variability of the speech tokens.

In both studies, one of the tasks consisted of attending to the whole interleaved sequence (i.e., grouping the two sequences together) regardless of the difference in F0 or in simulated positions between tokens, and indicating whether or not a repeated token was introduced. For good performance in this task, listeners should perceptually integrate the A and B sequences into a single stream, so that a repetition is heard within this stream, making it a measure of obligatory stream segregation (Micheyl and Oxenham, 2010). Since a difference in F0 would only influence the segregation of vowels, and since a spectral difference induced by a difference in elevation would only influence the segregation of the consonants, the analysis of the results focused on assessing whether consonants were separately grouped from vowels or whether they nevertheless remained grouped due to their occurrence in one syllable.

3. Stimuli and procedure

David et al. (2017a, b) used differences in spatial cues and F0, respectively, to study streaming of speech sounds. The stimuli used in both studies were naturally uttered pairs of voiceless fricative consonants and voiced vowels. They were recorded by a male native speaker of American English as a whole, so that they included a fricative part (the consonant), a formant transition part (the vocalic part still containing some consonant information) and a voiced part (the vowel). In the first study (David et al., 2017a), a set of 45 stimuli were recorded (five voiceless fricative consonants – [f], [s], [ϴ], [ʃ] and [h] – combined with nine vowels – [æ], [e], [iː], [I], [ʊ], [ɛ], [ʌ], [ɑ] and [uː]). In the second study (David et al., 2017b), 36 stimuli were recorded (four consonants – [f], [s], [ϴ] and [ʃ] – combined with the same nine vowels). Each single recorded stimulus will be referred to as speech token in the rest of the manuscript.

The stimuli had to be short enough to potentially produce obligatory stream segregation (van Noorden, 1975), but long enough to contain the information from both the consonant and the vowel. The duration of each speech token was therefore limited to 160 ms with 40-ms inter-token intervals, leading to an onset-to-onset time of 200 ms, which is close to the upper limit for observing obligatory stream segregation. The consonant and the vowel parts had approximatively the same length. The pitch contour of the tokens were flattened (110 Hz, unless specified otherwise) using Praat Software (Boersma and Weenink, 2017), where the stimuli were resynthesized using a pitch synchronous overlap-add technique (PSOLA).

In the first study, the ΔF0s tested were 0 semitones ($F0_A$ = 110 Hz, $F0_B$ = 110 Hz), 3 semitones (104 and 123 Hz), 5 semitones (98 and 131 Hz), 7 semitones (92 and 139 Hz), 9 semitones (87 and 147 Hz), and 13 semitones (78 and 165 Hz). In the second study, the differences in elevation angle separating the sequences were 0º, 10º, 50º and 70º.

In 50% of the trials, a repetition of a full token (consonant and vowel, “full repeat”) was presented. In 25% of the trials, a repetition of only the consonant was presented and in the last 25%, a repetition of only the vowel was presented. According to this paradigm, the Hits (Hs) correspond to the proportion of full repeats that were correctly reported and the False Alarms (FAs) correspond to the proportion of trial in which a repetition was reported when only a half-repeat was presented. Hence, it was possible to calculate separately the FAs for the consonant-only and the vowel-only repeats. The repeat, if present, occurred always at the penultimate position and the length of the sequence varied between 8 and 14 pairs of tokens. Figure 1 represent the tokens in the across-sequence task in the three conditions.

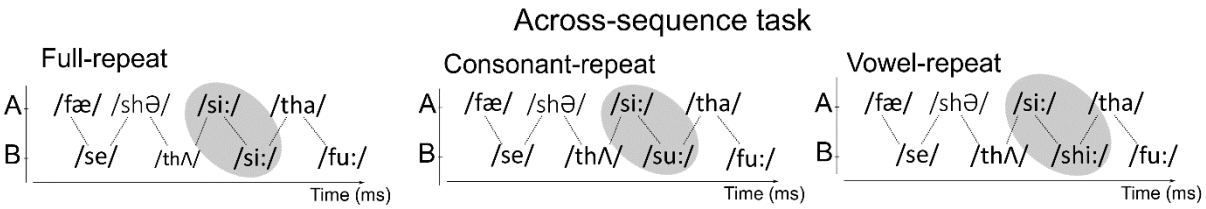

Fig.1: Schematic diagram of the tokens in the across-sequence task. From left to right the diagrams represent the full repeat, the consonant-repeat and the vowel-repeat.

4. Separate contributions of vowels and consonants to repetition detection

It was found that obligatory stream segregation could be observed based on a F0 difference (David et al., 2017b) and coloration differences in the median plane (David et al., 2017a). Fig. 2 represents the results of the first and second studies in terms of mean rates for the Hits and False Alarms.

In both cases, the Hit rate decreased, and thus segregation increased as the difference between the two sequences increased (a ΔF0 or a difference in spectral coloration in the median plane). There was no significant difference between the FAs due to a repeat of the consonant only or of the vowel only. Therefore, the authors then concluded that the listeners based their judgments on the whole CV token. This result is striking because the CV token seems to remain grouped, even when segregation occurs based on cues that affect only one part of the CV: F0 differences affect mostly the vowel (voiced) part, whereas coloration in the median plane, effective mainly at high frequencies, affects mostly the consonant part.

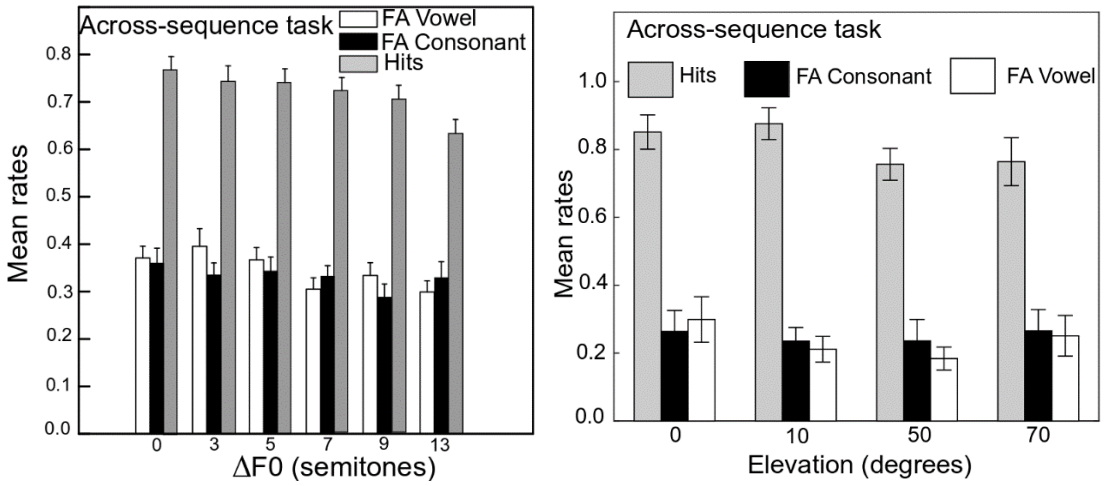

Fig. 2: Mean FA and Hit rates for the across-sequence task for the first study (left panel) and for the second study (right panel).

5. Discussion

The findings of the above studies suggest that in real life, when two sequences of sounds are perceptually segregated, all the syllables from one sequence are grouped together and segregated from the second sequence. From the studies discussed, it seems that the consonants and the vowels belonging to one sequence are never streamed apart. One possible interpretation of this result might be a top-down process. When learning a language, we acquire knowledge of what syllables and words exist. In the present studies, the CV tokens were common syllables used in the American English language and uttered by a human native speaker of American English. The listeners were also native speakers of American English. The knowledge of the language might have helped the listeners to group the tokens even when segregation occurred. To support this hypothesis, Jusczyket al. (1994) showed that infants learn the "rules" governing permissible sequences of phonemes in a given language allowing to compose words. Each language permits different sequences, for instance the syllable / zb / is not found in English but is common in Polish. Infants are able to use the transitions probabilities between adjacent tokens to detect words. They learn by being exposed to the right lexicon inherent to their native language (for a review, see Kuhl, 2004).

Another possible explanation might come from the formant transition. For example, Whalen (1991) showed that a whole fricative noise and the formant transition are needed for the listener to identify the fricative. Another example, taken from David et al. (2017b), showed that the voiced portion of a CV token does not carry enough information about the consonant to enable an accurate identification required in a streaming task (see experiment 3). It has also been shown that the recognition of monosyllabic speech sounds requires the adjacent formant transition (Lindblom and Studdert-Kennedy, 1967). Stachurski et al. (2015) investigated the importance of formant transition and continuity in F0 contours on the binding of speech sounds. They found that both cues help to keep the perceptual coherence of speech sounds. In addition, Cole and Scott (1973), in a task where listeners had to repeat the order of a sequence of speech sounds, found that vowel transitions help to preserve the temporal order of the sequences, so to preserve the coherence of natural speech. These findings concern mostly speech understanding. However, we can assume that stream segregation requires speech identification, and thus these results can be extended to streaming.

The two explanations, top-down process and formant transitions, are not exclusive. Indeed, Wagner et al. (2006) investigated the use of formant transition as a function of the spectral content of the fricatives present in the native language of the listeners. They tested five different languages: Dutch, English, German, Polish and Spanish. The participants had to recognize a fricative target in pseudo-words, the fricatives being preceded and followed by the vowels / a i u /. The results showed that depending on the native language, the listeners paid or did not pay attention to the formant transitions. Indeed, for the languages where there are no spectral similarities in the fricatives (e.g., Dutch and German), the listeners were not affected by a misleading formant transition. However, for the languages presenting spectral similarities in some fricatives (e.g., / f / and / th / in English and Spanish), the listeners appeared to pay attention to formant transitions. Further investigation would be

required to understand the binding of speech in languages where listeners do not pay attention to the formant transitions (e.g., Dutch or German).

When using synthesized vowels, Gaudrain and Grimault (2008) observed that different formants of the same vowel were segregated to form separated auditory streams. This outcome might be explained by the micro-modulations of frequency and amplitude present in the natural utterance, which is missing in the synthesized vowels. Besides, for naturally uttered sequences of vowels, where the spectral cues varied across time, it was found that segregation was not based on the spectral formant cues (see Gaudrain and Grimault, 2008; Gaudrain et al., 2007). Indeed, if segregation was based on vowels' spectral similarities, it would lead to a grouping unfavourable for a coherent speech perception (one stream of /a /, another stream of / i / etc… instead of different streams of meaningful syllables). The same is true for the organization of speech syllables like the CV tokens tested in the previously detailed studies. In this situation, the auditory system seems to be able to ignore the spectral variations inherent to speech, and thus grouping in speech may have rely on top-down information.

Acknowledgments

This work was supported by NIH grant R01 DC007657 and R01 DC016119 (AJO), Erasmus Mundus Auditory Cognitive Neuroscience travel award 22130341 (MD) and LabEX CeLyA ANR-10-LABX-0060/ANR-16-IDEX-0005 (ML, NG). We thank Steven van de Par for helpful discussions.

References

Blauert, J. (1997). *Spatial hearing : the psychophysics of human sound localization*. MIT Press.

Boersma, P., & Weenink, D. (2017). "Praat a system for doing phonetics by computer," [Computer program]. *Version 6.0.31, Last Retreived 21 August 2017 Form Http://Www.Praat.Org/.*

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sounds*. (MIT Press, Ed.). Cambridge.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.*, *25*, 975–979.

Cole, R. A., & Scott, B. (1973). Perception of temporal order in speech: the role of vowel transitions. *Can. J. Psychol.*, *27*(4), 441–449.

David, M., Lavandier, M., Grimault, N., & Oxenham, A. J. (2017a). Discrimination and streaming of speech sounds based on differences in interaural and spectral cues. *J. Acoust. Soc. Am.*, *142*(3), 1674–1685.

David, M., Lavandier, M., Grimault, N., & Oxenham, A. J. (2017b). Sequential stream segregation of voiced and unvoiced speech sounds based on fundamental frequency. *Hear. Res.*, *344*, 235–243.

Gaudrain, E., & Grimault, N. (2008). Streaming of vowel sequences based on fundamental frequency in a cochlear-implant simulation. *J. Acoust. Soc. Am.*, *124*(5), 3076–3087.

Gaudrain, E., Grimault, N., Healy, E. W., & Béra, J.-C. (2007). Effect of spectral smearing on the perceptual segregation of vowel sequences. *Hear. Res.*, *231*(1–2), 32–41.

Gockel, H., Carlyon, R. P., & Micheyl, C. (1999). Context dependence of fundamental-frequency discrimination: lateralized temporal fringes. *J. Acoust. Soc. Am.*, *106*(6), 3553–3563.

Hartmann, W. M., & Johnson, D. (1991). Stream Segregation and Peripheral Channeling. *Music Percept.: An Interdisciplinary Journal*, *9*(2), 155–183.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants′ Sensitivity to Phonotactic Patterns in the Native Language. *J. Mem. Lang.*, *33*(5), 630–645.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831–843.

Lindblom, B. E. F., & Studdert-Kennedy, M. (1967). On the Rôle of Formant Transitions in Vowel Recognition. *J. Acoust. Soc. Am.*, *42*(4), 830–843.

Micheyl, C., & Oxenham, A. J. (2010). Objective and Subjective Psychophysical Measures of Auditory Stream Integration and Segregation. *J. Assoc. Research in Otolaryngol.*, *11*(4), 709–724.

Miller, G. A. (1957). The masking of speech. *Psychol. Bull.*, *44*(2), 105–129.

Moore, B. C. J., & Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acusti. United Ac.*, *88*(3), 320–333.

Moore, B. C. J., & Gockel, H. E. (2012). Properties of auditory stream formation Properties of auditory stream formation. *Phil. Trans. R. Soc. B*, *367*, 919–931.

Sach, A. J., & Bailey, P. J. (2004). Some characteristics of auditory spatial attention revealed using rhythmic masking release. *Percept. & Psychophys.*, *66*(8), 1379–1387.

Stachurski, M., Summers, R. J., & Roberts, B. (2015). The verbal transformation effect and the perceptual organization of speech: Influence of formant transitions and F0-contour continuity. *Hear. Res.*, *323*, 22–31.

Stainsby, T. H., Fu, C., Flanagan, H. J., Waldman, S. K., & Moore, B. C. J. (2011). Sequential streaming due to manipulation of interaural time. *J. Acoust. Soc. Am.*, *130*(2), 904–914.

van Noorden, L. (1975). *Temporal Coherence in the Perception of Tone Sequences*. *Institute for Perception Research*. Eindhoven.

Vliegen, J., & Oxenham, a J. (1999). Sequential stream segregation in the absence of spectral cues. *J. Acoust. Soc. Am.*, *105*, 339–346.

Wagner, A., Ernestus, M., & Cutler, A. (2006). Formant transitions in fricative identification: The role of native fricative inventory. *J. Acoust. Soc. Am.*, *120*(4), 2267–2277.

Whalen, D. H. (1991). Perception of the English /s/–/ʃ/ distinction relies on fricative noises and transitions, not on brief spectral slices. *J. Acoust. Soc. Am.*, *90*(4), 1776–1785.

Wightman, F. L., & Kistler, D. J. (1992). The dominant role of low-frequency interaural time differences in sound localization. *J. Acoust. Soc. Am.*, *91*(3), 1648–1661.