Cognitive and methodological considerations on the effects of musical expertise on speech segmentation
Clément François, Barbara Tillmann, Daniele Schön

To cite this version:


HAL Id: hal-02062407
https://hal.archives-ouvertes.fr/hal-02062407
Submitted on 8 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Cognitive and methodological considerations on the effects of musical expertise on speech segmentation

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Annals of the New York Academy of Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>annals-9999-732</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Conference papers</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>31-Oct-2011</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>François, Clément; INCM, CNRS Tillmann, Barbara; Lyon Neuroscience Research Center CRNL, CNRS Schön, Daniele; BDI, INSERM</td>
</tr>
<tr>
<td>Keywords:</td>
<td>statistical learning, language, music</td>
</tr>
</tbody>
</table>
Cognitive and methodological considerations on the effects of musical expertise on speech segmentation

Clément François\textsuperscript{a}, Barbara Tillmann\textsuperscript{b}, Daniele Schön\textsuperscript{c}

\textsuperscript{a} INCM-CNRS UMR6193 & Aix-Marseille University, France
\textsuperscript{b} CNRS UMR5292, INSERM U1028, Lyon Neuroscience Research Center, Auditory Cognition and Psychoacoustics Team, Lyon, France & Université de Lyon, France
\textsuperscript{c} Brain Dynamics Institute, INSERM & Aix-Marseille University, France

Abstract

Both speech and music are constituted by sequences of sound elements that unfold in time, and require listeners to engage cognitive functions such as sequencing, attention and memory. We recently ran a set of experiments aiming at testing the effect of musical expertise on a rather high cognitive function: speech segmentation. Here we will present the main concepts underlying the investigation of speech segmentation as well as its link to music and musical expertise. Interestingly our results seem to show that musical training and expertise have effects on brain plasticity that may go beyond primary regions. Moreover, to facilitate and improve future research in this domain, we will here describe several delicate methodological precautions that need to be taken into account (e.g., the choice of stimuli, participants, data analyses). Finally, we will give some possible future directions to better understand the impact that music may have on speech processing.
An introduction to statistical learning

Learning the mother tongue or a second language is a rather long process that goes through several dependent phases. Because word boundaries are not systematically flagged by acoustic cues such as stresses or pauses, one important step in language learning is the ability to extract words that unfold in time. Many studies inspired from the seminal work of Saffran and colleagues showed the importance of the statistical structure of the speech stream for an efficient segmentation\textsuperscript{1-3}. Indeed, in a speech stream, within-word syllables tend to be associated more often than between-words syllables. The importance of these statistics (called conditional or transitional probabilities) has been shown in adults, infants and neonates\textsuperscript{4-7}. The experimental paradigm consists of a familiarisation (learning) phase followed by a test. During the familiarisation phase, participants listen to several minutes of a statistically structured, continuous flow of artificial syllables without any acoustic cues at word boundaries. The test phase depends on the participant population (infants, adults). In the case of adult participants, the test is a two-alternative forced choice procedure (AFC) and participants have to choose, in each trial, between a word that was part of the language and a word built with similar syllables, but that was not part of the language (henceforth partial word). Above-chance performance suggests participants’ ability to segment the auditory (speech) stream on the basis of transitional probabilities. This ability has been also demonstrated for non-linguistic auditory sequences: such as sequences of pitches\textsuperscript{4,8,9} or of instrumental timbres\textsuperscript{10}, thus suggesting that this type of learning is not specific to speech. Finally, this ability has also been shown for sequences of visual stimuli (shapes or movements) and sequences of tactile stimulations\textsuperscript{11} suggesting that this type of learning is at work in several modalities\textsuperscript{12-14}.

Song is particularly well-suited to have a better understanding of the relations between music and language processing. Indeed, the segmentation of linguistic and non-linguistic inputs has been studied separately (with different tasks and different participants), thus rendering difficult to draw any
clear conclusion on the non-specificity of the learning processes for language and music. Recently, we compared speech segmentation of an artificial sung language (with speech and music combined in the same signal) to speech segmentation of a spoken language. Speech segmentation was better when sung than when spoken, possibly due to structural and motivational properties of music. In François and Schön, participants (nonmusicians) were exposed to the sung speech stream, but then tested separately on linguistic and musical dimensions of the sung language (both behavioural and EEG responses were recorded at test). Performance was above chance for the linguistic test, but at chance level in the musical test. By contrast to this behavioural measure, the analysis of the event-related potentials (ERPs) revealed in both linguistic and musical tests a similar fronto-central late negative component that was larger for non familiar (partial words/sequences) than for familiar items (words/sequences). In agreement with previous findings, we interpreted this late negative component as an index for the search of memory traces that have been shaped during learning (i.e., in the familiarisation phase). In François and Schön, we compared sung speech segmentation in nonmusicians and professional musicians. While the behavioural results did not show a clear effect of expertise, ERP data showed a larger late negative component for musicians than for nonmusicians, in both language and musical tests.

Finally, to show that music training was the cause of this difference (and not prior differences of the participants in the two groups), François et al. conducted a longitudinal study spanning over two years using a test-training-retest procedure with a pseudo-random assignment of children to two different artistic training programs. Children followed a training on either music or painting, and were tested on their ability to extract words from a continuous flow of syllables (the linguistic test). Both behavioural and electrophysiological measures showed a greater improvement in speech segmentation across testing sessions in the music group compared to the painting group. Taken together, these findings point to a benefit of musical expertise and musical training for both speech and music.
Why would speech segmentation benefit from musical expertise?

In 2007, Hickok and Poeppel proposed a dual route model of speech processing, which includes dorsal and ventral pathways. The dorsal pathway acts as a sensori-motor interface aiming at mapping the phonologico-acoustic representations of speech sounds to articulatory representations. The ventral pathway acts as a lexico-conceptual interface aiming at mapping the phonologico-acoustic representations to lexico-conceptual information. Based on a set of studies combining behavioural, EEG and fMRI techniques, Rodriguez-Fornells et al. have recently adapted and completed this model for speech segmentation. They described a large cortical network involved in speech segmentation comprising the posterior part of the superior temporal gyrus (STG) and the premotor cortex (PMC) connected via the arcuate fasciculus.

Within this framework, several non-exclusive hypotheses can explain the differences in speech segmentation between musicians and nonmusicians. First, musicians might benefit from more efficient sound encoding at subcortical and cortical levels. Indeed, previous studies have shown functional differences between musicians’ and nonmusicians’ encoding of musical and linguistic sounds. In our experiment, we also found differences due to expertise in rather early auditory components, such as the N1, the P2 and the mismatch negativity, all possibly generated in the primary auditory cortex and the planum temporale. Thus, early differences in the functioning of the brainstem and the auditory cortex might explain the musical expertise effect, notably with musical training leading to a reorganization of auditory neurons along the auditory dorsal pathway, thus facilitating the subsequent processing steps in the STG and the left inferior frontal gyrus (IFG).

Second, an alternative explanation could be that musicians may have a more developed and efficient dorsal pathway than nonmusicians. Its implication for statistical learning has been suggested...
by a recent study demonstrating that white matter integrity in the vicinity of the left IFG predicted performance in artificial grammar learning\textsuperscript{25}. Similarly, in a speech segmentation experiment combining fMRI and EEG measures, Cunillera et al.\textsuperscript{26} observed a fronto-central late negative component (at word onsets) that increased over the learning period and that had its generator around the PMC and the Left IFG. Taken together with the observation of increased grey matter density and volume in the left IFG for musicians in comparison to nonmusicians\textsuperscript{27}, these results suggest that musicians’ advantage in speech segmentation might be related to enhanced involvement of the pre-motor brain areas in comparison to nonmusicians.

Third, we might hypothesize that musical training improves the connectivity between these two subsystems. This hypothesis can be integrated in the recent proposition that sound acts as a scaffolding framework for cognitive sequencing\textsuperscript{28}, also supporting how to process and interpret sequential and temporal information in the environment. It is suggested that sound and speech processing has an additional unspecified influence on the development of general cognitive sequencing abilities (also in other modalities). While Conway et al.\textsuperscript{28} find support for their hypothesis in the consequences of auditory deprivation on domain-general sequencing (i.e., impaired implicit learning for visual, non-linguistic regularities in deaf children), additional support can be found in research investigating the effect of musical expertise, that is increased training on sound analyses and enhanced sound exposure should lead to improved sequencing also in non-musical tasks (such as speech segmentation).

Methodological considerations

(i) The comparison of musicians and nonmusicians

When comparing the performance of musicians and nonmusicians, one has to keep in mind some general difficulties and restrictions to conduct proper experiments testing for the effect of musical practice. Indeed, most studies comparing musicians and nonmusicians use a cross-sectional approach: a
group with 10 years of musical practice is compared to a group without musical practice possibly comparable in age, handedness, sex and level of education. These studies are very instructive, but always fall in the criticism of non-controlled pre-existing genetic or other factors explaining the observed between-group differences. These issues can be ruled out with a longitudinal approach testing naïve (with respect to music) participants enrolled in a test-training-retest procedure, while still controlling for several standardized neuropsychological tests and several socio-economic variables. After the first testing session, this specific procedure requires a pseudo-random assignment of participants to either a musical group or a control group. This procedure also encounters some difficulties, like the critical choice of the activity proposed to the control group and the repetition of the tests which renders the test explicit. This is particularly delicate when interpreting speech segmentation as an implicit learning process (cf. Section iii).

(ii) The choice of the instructions

The instructions before the learning phase can become, in our opinion, important because they determine to which extent learning may take place implicitly or not. For instance, in some studies, participants are explicitly told about the presentation of an invented “non-sense” language, or even to look for the words embedded in the speech stream (e.g., Saffran et al.29: “figure out where the words begin and end”). Another study present the isolated words (visually) before the familiarisation phase in order to maximize learning17. By contrast several studies, like ours, do not give any explicit information before the familiarisation phase: participants are asked to carefully listen to a continuous stream of syllables for several minutes.

(iii) Measuring learning: from explicit to implicit measures

Another methodological difficulty refers to the implicit nature of the tests and, of course, the
implicit nature of the learning process. Studies investigating speech segmentation mostly refer to “statistical learning” rather than “implicit learning”\textsuperscript{30}, which leads to the absence of testing whether the learned knowledge is implicit. The domain thus adapts a definition of implicit learning as the incidental nature of the acquisition process and without the intention to learn (at least for some of the instructions, see above), rather than extending it to the implicitness of the acquired knowledge\textsuperscript{31,32}.

The standard behavioural test used in this domain can be criticized for theoretical purposes in the field of implicit learning. The two-alternative forced choice procedure requires participants to make an explicit judgement on the two presented items without feedback (“indicate which of the two strings sounded more like a word from the language you have heard before”). However, the representations of the segmented items might be weak and might vanish rapidly in time, probably also due to the interference caused by the presentation of non-familiar items. Thus, there is a need to promote implicit measures, like ERPs, which do not necessary require an overt behavioural response (e.g., analysis of ERPs during the familiarisation phase). The priming paradigm could also be adapted to the testing of speech segmentation as it does not require telling participants about the structure of the material and as it has already been successfully applied to artificial grammar learning\textsuperscript{33}. Similarly, the Serial Reaction Time procedure can be also used as a promising implicit measure, as introduced by Hunt and Aslin\textsuperscript{12} for the visual modality with 3-element units: Over exposure, response times decreased for elements within the units compared to response times for elements crossing unit-boundaries.

Another relevant point to consider is the use of repeated tests in within-subject designs, such as for test-training-retest designs or designs comparing different types of languages or studying the transfer from one language structure to another (see, for example, Omigie & Stewart\textsuperscript{34}). Indeed, when interpreting results and comparing them to other results reported in the literature, one needs to be aware of the fact that once tested with a typical AFC task, learning of a successive language structure will be even less implicit, insofar as subjects are then aware that the stream contains words and possibly of the
structure of the stream (e.g., trisyllabic words). One possibility to at least blur the cues for potential strategies gained in the first test phase and then orienting the perception of the second exposure phase is the use of test items of different lengths (bi- and trisyllabic items), even though only trisyllabic words are relevant.

(iv) The control of the stimuli (units and stream)

When preparing an artificial language, great care must be taken about the choice of the elements, the definition of the units and their chaining in the stream. The first choice is related to the choice of the phonemes. For instance, a language containing consonants t, d, b, p and vowels i, u, will be more difficult to segment than a language containing t, s, m, p and a, o, due to the greater phonetic proximity in the former compared to the latter. Also, when planning to acquire EEG, one may want to use consonants with a short and similar attack time (e.g., plosives) in order to have clearer ERPs (N1-P2 complex) to the onset of syllables/words.

Another important parameter is the number of words to be used to build the stream (typically between 4 and 6). Because the stream is typically generated by a pseudo-random concatenation of the words (no repetition of the same word twice in row), using a very small number of words has a direct impact of the transitional probabilities at word boundaries. In the extreme case of using two words (e.g., bada-tibu), the transitional probability at word boundaries (da-ti, bu-ba) would be identical to the transitional probabilities between the two words. This might be particularly relevant, when using partial words in the test phase (in particular for the 2-AFC test), because, when using a few words, TPs for the partial words will be very close to TP for words (identical in the case of 2 words). Thus, one may want to calculate TPs for words AND partial words before running the experiment. Also note, that while most studies used only trisyllabic words, some studies started to use words of different lengths aiming to get closer to real-world language learning situations.
Another parameter that needs to be taken into account is the ratio between the duration of the familiarisation phase and the number of words, weighted by the number of syllables building a word. These choices can vary considerably across the published experiments, going from a 2-minute stream using 4 bisyllabic words (in babies\(^1\)) up to a 21-minute stream using 6 trisyllabic words (in adults\(^29\)).

One may also note that while behavioural experiments often use rather long and boring familiarisation durations, EEG analysis seems to show that learning may take place in the very first minutes\(^{26,37}\). This suggests that the 2-AFC might not be a sensitive measure of implicit learning, as already discussed above (cf. Section iii).

Another decision to be taken concerns the definition of the 2-AFC test phase: participants have to choose between one item that is a word and the other item that is not. For the latter one, three types are commonly used: non-words for which syllables are arranged in an order that has never occurred in the familiarisation phase; partial words for which two of the three syllables have appeared in that order in the familiarisation phase, but the association with the third syllable has never been heard; and partial words that have been heard in the familiarisation phase, thus containing the boundary between two words (e.g., da-ti in the 2-words language example given above). Of course, the choice of one of the three test item types influences the interpretation in terms of learning processes. When using non-words, participants may simply rely on the detection of a new transition between two syllables, while the use of partial words discards this possibility and is thus more informative for statistical learning (see\(^{38}\) for a more complete discussion on this issue).

For the speech material preparation, another delicate point that is rarely addressed in the literature is the synthesis procedure used to generate the stream. Indeed, because acoustic cues need to be controlled, a voice synthesizer is used to generate speech (even though cross-splicing techniques might represent an alternative method). However, speech synthesis is far from being a simple affair and it is thus important to carefully check the acoustic features of the generated stream. This is even more
important when using a sung language, insofar as there might be unexpected interactions between
vowels and pitch, resulting in perceptual accents or lengthening as well as in more or less clearly
pronounced consonants.

The importance of a careful selection in the material construction is not restricted to verbal
material only, but also affects non-verbal segmentation tasks. In the case of music, for instance tone
triplets may pop out within the stream (reinforcing or obscuring boundaries). When tone sequences (or
sung sequences) are used, special care must be taken to the choice of the pitches (and pitch classes),
their potential link with musical structures (e.g., tonality), but also the defined interval sizes and
interval directions between adjacent tones (e.g., jumps, reversals) either within units or across unit
boundaries (see also\textsuperscript{4}). When musical timbres are used, care must be taken regarding acoustic similarity
(or more generally surface features) for the definition of units and boundaries (see\textsuperscript{10}). It is worth
underlining that these indications only concern the choice of the various features in relation to the
definition of units and their boundaries in the stream. Nonetheless, implicit learning can be studied with
materials varying in distance and similarity with real-life materials (e.g., Loui et al., showing artificial
grammar learning with a new musical scale\textsuperscript{39}).

Regarding the selection of elements and units, some experimental controls have been proposed
(though not systematically applied yet) to ensure that above chance performance is not due to some
preference bias of syllable (or tone) combinations in statistical units or other unrelated influences. 1) Saffran et al.\textsuperscript{4} proposed to define two language systems (L1, L2) that were built in such a way that L1
consists of partial units of L2, and vice versa. This construction allows using the same test phase for
two participant groups (having been exposed to one or the other language), and thus avoiding a
confound between learning and other influences (perceptual biases, preferences, etc...). 2) Reber and
Perruchet\textsuperscript{40} suggested to use multiple implementations: all languages have the same statistical
structures, but are instantiated by different syllable attributions (see also\textsuperscript{41}).
(v) Data analyses

In several statistical learning studies, t-tests are used to make statistical inferences on whether learning has taken place or not. In particular, group performance is compared to chance level. While the two sample t-test is not appropriate because a “chance” sample has no variance, one should use a one sample t-test, setting up a normal distribution with a mean specified by the null hypothesis (here 0.5). Also, non-parametric tests, such as the one-sample Wilcoxon Signed Rank test, might be more appropriate. Another (non-exclusive) approach is to have either a between-subject design directly comparing results of two or more samples/populations or a within-subject design directly comparing, for instance, learning of the linguistic dimension and the musical dimension, or learning before and after music training (however, see above for comments on repeated-measure designs). Finally, a solution at the individual subject level might be to use binomial tests to estimate the exact probability for each individual. Then, one may run second-order statistics on the p-values, although this might lead to very conservative test (with 36 trials the individual threshold for significance would be 24 correct responses, p = .0326).

Another type of analysis consists in taking into account the differences between items. Because each item has average transitional probabilities between syllables (TPs), it is possible to test whether items with high TPs have been learned better than items with low TPs. At this aim, one may either run a non-parametric ANOVA with items as a factor, or contrast directly high versus low TPs items. These analyses should point out the importance of TPs, by showing poor performance for low TP items and good performance for high TP items\(^ {29} \).

To better define the role of the TPs in the learning process one could also take into account the distinctiveness between the two items in each trial. This is of specific interest when using partial words that have been heard during the familiarisation phase (i.e., TPs>0) and that by consequence may also
have a sort of pre-lexical representation competing with words. The underlying idea is that trials with a high TP contrast between words and partial words (i.e., very different TPs) should be easier and thus show better performance than trials with a low TP contrast (i.e., very close TPs).

Finally, another interesting way to model data is, in our view, taking into account time. As we discussed above, the test is rather explicit and can lead to interference or learning effects along the testing session (typically 36 trials). Thus, modelling time in the statistical analyses by comparing performance in the first half of the test to performances in the second half of the test can be very instructive in this respect. For instance, Rohrmeier et al.\textsuperscript{42} used an artificial grammar paradigm with sequences of tones and showed increasing performance along the testing session for the untrained group while the trained group showed decreasing performance along the test.

**Speech segmentation and musical expertise: perspectives and future directions**

This last section presents possible future directions in the field of musical training and/or practice-shaped brain plasticity and statistical learning, keeping in mind that the most important point, in our view, is understanding what aspects of music training (and its consequences on the brain) might contribute to beneficial effects for speech segmentation and to what extent. Several non-exclusive directions might be interesting in this respect. For instance, aiming to determine the relevant processes, one could study the effect of different types of musical trainings, such as rhythmic training versus pitch/tonal training, by comparing, for example, a group of drummers to a group of singers.

Another promising perspective will be investigating the relation between perceptual and productive musical skills and their interaction with speech segmentation processes. For this aim, psychophysical measures of perceptual and productive skills could be correlated with statistical learning performance as well as with EEG or fMRI markers of speech segmentation collected on the same participants. Furthermore, because musicians' advantage could be in part explained by a better
pre-attentive sound processing, other experiments could record both pre-attentive and attentive subcortical (e.g., frequency-following responses) and cortical auditory responses (e.g., MMN and P300) and correlate these measures with behavioural performance and psychophysical tests.

Another possible direction is to test interactions between musical expertise and different acoustic and linguistic properties of the speech segmentation input. For instance, it has been shown that lexical stress (including changes in pitch and timing) and subliminal gaps at word boundaries facilitate speech segmentation (e.g.,\textsuperscript{37,42,43}). Thus, it could be interesting to see to what extent musical practice modifies the sensitivity to these acoustic cues on both pitch and time dimensions.

On a more linguistic side, Tyler and Cutler\textsuperscript{36} compared English, French and Dutch speakers in an artificial language learning paradigm and reported an interaction of participants’ mother tongue and the influence of vowel lengthening and pitch movement on speech segmentation performance. Similarly, a speech stream respecting the native phonotactic rules results in better segmentation than a “foreign” speech stream\textsuperscript{44}. These manipulations of the stream could shed new light onto the extent to which musical practice may affect language-universals and/or language-specific knowledge.

Finally, a very promising direction will be to investigate learning processes per se, and to study the familiarisation phase using electrophysiological measures (EEG/MEG). This approach has the double advantage of an implicit measure of learning as well as of giving access to the learning curve over time. Interestingly, some studies have reported different learning curves as a function of behavioural performance, contrasting EEGs of high vs. low learners\textsuperscript{8,26,37,45}. For the present aim, advances in signal-processing methods, such as time-frequency analyses\textsuperscript{46} and Frequency-tagging analyses\textsuperscript{47}, will allow for a better understanding of the cortical processes participating in speech segmentation and will probably turn out to be highly informative when applied to investigate the bases of the differences between musicians and nonmusicians in speech segmentation.
Acknowledgments

Preparation of this paper was supported by a grant from the French National Research Agency to D. Schön and B. Tillmann (ANR Blanc DMBB #NT09_520631). C. François was PhD student supported by the ANR-Neuro (#024-01).

References


http://www.nyas.org/forthcoming


