Intra-speaker phonetic variation in read speech: comparison with inter-speaker variability in a controlled population

Nicolas Audibert¹, Cécile Fougeron¹

¹Laboratoire de Phonétique et Phonologie, UMR7018 CNRS/University Sorbonne-Nouvelle, Paris, France
{nicolas.audibert; cecile.fougeron}@sorbonne-nouvelle.fr

Abstract
Our knowledge of speech has historically been built on data averaged across speakers or comparing different speakers. We therefore know little about the variability of speech produced by the same speaker: to what extent does it vary from one repetition to another and in what dimensions? In this study, we document the stability of speech and voice characteristics in 9 French speakers, on the reading of two texts recorded over ten sessions on different days over a two-month period. 21 features related to temporal, spectral, F0 and harmonicity aspects as well as their modulation between consecutive chunks are studied. The stability of these features between sessions is evaluated in comparison with their variability between speakers. Results show that short-term variability of energy in the 0-1kHz band, mean F0 and the slope of the LTAS vary the most between sessions for a given speaker, and are also among the speech and voice features that vary the most between speakers in this small cohort, while modulation features between consecutive chunks remain more stable across sessions.

Index Terms: intra-speaker variability, French, voice and speech dimensions, cross-session stability

1. Introduction
Our knowledge of speech has historically been built on data comparing different speakers or on data averaged across speakers. Consequently, little is known about the variability of speech within an individual. Yet we know that speakers adapt to linguistic and speech contexts, and modify their speech according to their emotional or biological state, for example. However, it is not clear to what extent speakers vary from one recording session to another, on which speech features, and to what extent recordings collected days, months, or years apart are "naturally" variable.

While variability in speech is ubiquitous, it is also, at least in part, structured and governed by rules. Therefore, documenting variation in speech and explaining its origins is at the heart of phonetics research. Two types of variability can be distinguished according to their origins and characteristics: inter-speaker variability and intra-speaker variability [1].

On the one hand, inter-speaker variability is rooted in speaker-specific physiological or anatomical traits, as well as demographic, regional or social factors [2]. This inter-speaker variability carries indexical information about the speaker: this indexicality is disseminated over a cluster of speech features and it is usually this cluster that discriminates one speaker (or group of speakers) from another. Moreover, while speech conditions have been considered for a long time by researchers in voice identification and forensic phonetics as a factor of intra-speaker variability that may impair the ability to recognize a speaker’s voice (see e.g. [3], or Brunner [4] for a review), the influence of speech conditions on inter-speaker variability remains understudied.

On the other hand, intra-speaker variability comes from a variety of sources (see e.g., [5] for a review). Among these, changes in phonetic realization of phonemes, due to phonetic or prosodic contexts, have been the most documented and this intra-speaker variability is easier to model. This is not the case for other sources of intra-speaker variability which are due to the fact that individuals adapt their speech to the speech context, to its formality, to the social relationship they have with their interlocutors, to the specific needs of the listeners (in a noisy context for example, or if they share common knowledge on the subject or not), etc. These different factors, often grouped under the notion of "style effects", mainly influence the degree of articulatory precision (or hyper/hypo articulation) and the speech rate adopted by the speaker. A third type of factor affecting the way a speaker speaks is related to the speaker's affective and cognitive state, related for instance to emotion, attention or fatigue (see among others [6], [7], [8]).

Although recognized and documented in some way, changes in speech resulting from speech context or speaker state are difficult to delineate and manipulate experimentally (see for example the work on “clear speech” by Scarborough & Zellou [9]). Moreover, our knowledge of speaker-internal variation is further complicated by the fact that all of these effects are not static but change dynamically over the course of an utterance [10]. Indeed, environmental conditions, speech content, prosodic properties affecting pronunciation, as well as fatigue, emotion, arousal, or attention vary from moment to moment.

Therefore, a common and repeated belief in language science is that a unit of speech will never be pronounced the same way twice by the same individual. However, little is known about how much speech diverges from one iteration to the next, in what aspects, how much granularity is needed to measure these variations, or what factors contribute to the variability from one iteration to the next. Heald & Nusbaum [11], for example, explored variability in isolated vowel production for eight speakers across 9 repetitions collected either on the same day at three different times or on different days. They found greater variability in vowel acoustic features in different sessions on the same day than between days, and proposed a broad set of possible causes for this phenomenon.

In forensics, where an individual's recordings must be compared, these issues seem crucial. Indeed, as outlined by Rose [12], speech samples compared in forensic applications are never contemporaneous. Results obtained by Rhodes [13] show that performance of voice identification by an automatic system
decreases with longer delays between recordings of a same speaker on controlled material (7 to 28 years in his dataset), in line with changes in formant frequencies, though with variations between speakers. However, to the best of our knowledge, we have no scientific criteria concerning the validity of a comparison between recordings spaced only days, weeks, or months apart and on the conditions allowing such comparisons.

As part of a larger research project on intra- and inter-individual variability in databases collected at the LPP on the same speakers with different inter-session delays (PATAFreq and PATAFreq databases), the study presented here focuses on the stability of speech features read over sessions spaced closely together in time. In order to make sense of this inter-session variability, it is compared to the variability of these speech features between the different speakers.

2. Methods

2.1. Speech material

The speech data used for this study come from the PATAFreq recording database which was collected during the first lockdown due to the Covid-19 pandemic in 2020 by 9 speakers, including 6 women, all native French speakers, close to the speech research community, and participating in the PATATRA longitudinal recording database coordinated by the LPP lab. These 9 speakers recorded themselves regularly, with a goal of about 10 recordings over about a month and a half. Five speakers reached 10 sessions, 3 speakers reached 8 sessions, and one speaker who encountered technical problems could only complete 6 sessions. Speakers were asked to record regularly, with a minimum of 24 hours between sessions and over a maximum period of two months, and were encouraged to change the time of day, but family constraints during lockdown did not always allow this. All speakers were equipped with sound cards and head-mounted professional microphones and were required to record in a quiet environment, always in the same room if possible, and with the same settings (incl. distance mouth-microphone). Any unexpected changes or environmental noise had to be reported in a questionnaire to be filled in after each recording. To tackle the self-assessment on a 4-level scale of fatigue, emotional state, vocal fatigue and frequency of voice use on the day of the recording.

Seven of the nine speakers drove their recordings using a dedicated application developed in Python (a modified version of MonPaGe, [14]) that managed the presentation of instructions, recording, and organization of sound files. Two speakers were unable to use this application and made their recordings with Praat [22], relying on a version of the instructions displayed as a slideshow.

Among the 7 tasks of the PATAFreq protocol, we study here only the task of reading two texts (text 1 and 2) which were manually segmented into chunks, so as to obtain about ten syllables per chunk. In addition to the segmentation into chunks, all the perceived silent pauses were delimited, the pauses between two chunks being by convention attached to the first of the two.

2.2. Acoustic features

The 11 acoustic measures summarized in Table 1 were extracted with Praat [22] on each of the previously segmented chunks, using a custom script. With the exception of the measure of intra-chunk variability of the harmonic-to-noise ratio (sd(HNR)), these measures were also derived to obtain a measure of modulation between consecutive chunks, denoted as ”d(X)” (for instance the variable d(articRate) corresponds to the chunk-to-chunk modulation of the articulatory rate), and normalized as follows: \[ \text{chunk}_{i} \text{ } \text{chunk}_{i+1} \text{ } \text{(chunk}_{i} \text{ } \text{chunk}_{i+1})/2 \]. As a result, we get a total of 21 descriptors on each of the chunks (except for the last chunk of the texts).

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Nom</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>articRate</td>
<td>Speech rate, pauses excluded (phones/sec.)</td>
</tr>
<tr>
<td></td>
<td>pauseTime</td>
<td>Cumulated time of silent pauses (seconds)</td>
</tr>
<tr>
<td>Pitch</td>
<td>meanF0</td>
<td>Mean F0 level (Hz)</td>
</tr>
<tr>
<td></td>
<td>F0range</td>
<td>F0 spread in the 10%-90% percentile range (semitones)</td>
</tr>
<tr>
<td>Harmonicity</td>
<td>meanHNR</td>
<td>Mean harmonics to noise ratio (dB)</td>
</tr>
<tr>
<td></td>
<td>sd(HNR)</td>
<td>Intra-chunk variation of HNR (dB)</td>
</tr>
<tr>
<td>Spectrum/</td>
<td>LTASslope</td>
<td>Slope between 0-1kHz and 1-4kHz frequency bands of the long-term average spectrum on voiced parts (dB)</td>
</tr>
<tr>
<td>Voice</td>
<td>sd(energy0-1kHz)</td>
<td>Variability of intensity (dB) on 4 freq. bands: 0-1kHz, 1-2.5kHz, 2.5-4kHz (articulation of segments), and 4-8kHz (lesser influence of segments)</td>
</tr>
<tr>
<td></td>
<td>sd(energy1-2.5kHz)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sd(energy2.5-4kHz)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sd(energy 4-8kHz)</td>
<td></td>
</tr>
</tbody>
</table>

2.3. Analyses

The relative importance of the 21 measures selected for distinguishing between sessions produced by the same speaker was evaluated using a variable selection procedure based on random forest classification and implemented in the R package Boruta [15]. In this procedure, the evaluation of the importance of each variable is based on the estimation of the loss of precision in the classification task induced by the random permutation of values between elements to be classified. Each variable is attributed an importance score allowing to prioritize them and eventually to discard the variables not resulting in a better classification than the random reference. In addition to classical random forest methods, the general principle of the Boruta algorithm is to use shadow variables obtained by applying random permutations to each predictor as baseline, the shadow variable with the highest importance score being retained as a reference. Importance scores are determined over multiple runs, resulting in a distribution of importance score values for each variable. The efficiency of the Boruta algorithm was assessed by [16], and this method has already been applied to phonetic data by [17].
In our case, this procedure was applied separately to the data of each speaker for the classification in sessions, and on the whole dataset for the classification in speakers, over 100 runs. We do not consider here the classification in texts. In order to allow a direct comparison between classification tasks involving a non-homogeneous number of classes and observations, we use as a metric the rank of each variable after sorting them by importance score.

3. Results

Figure 1 presents the descriptors ordered by rank according to their importance for the separation into sessions for each speaker, and for the separation into speakers. The distributions of the descriptors discussed below are illustrated in Figure 2, for each consecutive session and each speaker.

Among the descriptors that stand out among the 5 variables estimated to be most important by the Boruta classification for separation into sessions (at the top of Figure 1 and Figure 2) we find: meanF0 for all speakers except F05 (for whom it appears at rank 7), sd(energy(0-1kHz)) for all except F02 (at rank 7), LTASslope for all except F04 and H02 (at rank 6 and 7 respectively) and meanHNR for all except F05 (at rank 7) and H01 (at rank 12). All these descriptors which characterize low frequency aspects, supposedly related to the voice, are therefore quite unstable from one session to another and/or within the same session (i.e. between chunks). The comparison of the weights of these descriptors with those estimated to be the most important for the classification into speakers (Figure 1) reveals the variables that were found to be among the most important for the separation into sessions are also those that differentiate our 9 speakers the most, with a rather high Spearman correlation between the rank for the separation in speakers and the mean rank for the separation in sessions for each speaker (ρ=.81). Indeed, meanF0, LTASslope and sd(energy0-1kHz) are among the top-5 descriptors useful for speaker separation. These descriptors distinguish between sessions of the same speaker but also between speakers. For instance, we can see on Figure 2 that according to the session, the average F0 of speaker F02 is either closer to that of speaker F05 (sessions 3-6-9-10), or closer to that of F03 on other sessions.

On the selection of descriptors presented in Figure 2, it is also clear that some speakers are quite variable from one session to another while others are remarkably stable, like F03 and H01. As also illustrated in Figure 2, the stability/variability of the articulatory rate between sessions depends on the speaker: this descriptor ranks in the top-10 variables for session separation for F01, F03, F04, F05 and H01, while it ranks much lower for F02, F07, H02 and H03. Among these 4 speakers, the patterns are different: for F02 and H02 the articulatory rate is particularly stable from one session to another and also shows a restricted range of variation within the same session (between chunks). F07 and H03, on the other hand, show much more variation in their productions within the same session and between sessions. These individual characteristics of articulatory rate explain why this descriptor reaches the top 5 of speaker separation on the cohort.

Frequency fluctuations in the 4-8kHz band (sd(energy(4-8kHz)), also shown in Figure 2, are as expected among the top 5 descriptors for speaker separation. However, we also notice that they are also variable between sessions and present a significant range of variation within sessions: for all speakers except F03, this descriptor thus appears in the top-10 of the separation into sessions. Its variation between chunks within the same session suggests that in this frequency band, we capture more variations related to the segmental content than expected.

Chunk-to-chunk modulation of features along the text reading (d(X)) have little weight in the separation between sessions of the same speaker (variables listed at the bottom in Figure 1 and Figure 2). Among these, the chunk-to-chunk modulation of the energy fluctuation in the mid-frequencies (d(sd(energy(1-2.5kHz)) or d(sd(energy(2.5-4kHz)) depending on the speakers) are found among the 5 most stable measures between sessions for all speakers. If we consider that on these frequency bands, we capture spectral variations related to the dynamics of the articulation of sounds, we can interpret this result as reflecting the stability of this articulatory chunk-to-chunk dynamics. The chunk-to-chunk modulation in high frequencies (4-8kHz) also appears among the most stable descriptors for 7 speakers out of 9. The HNR modulation (d(meanHNR), not shown here) is also stable for all speakers except F05 and F07, and could reflect a higher variability of voice quality along the reading, possibly linked to vocal fatigue. However, this interpretation remains to be evaluated through measures of the evolution over time in the same session.

4. Discussion and conclusion

While it has been documented for a very long time, speech variability within the same individual is rarely studied. In contrast to other studies conducted either on a large number of speakers but on a small number of repetitions (e.g., Keating et al. [18]: 200 speakers, 12 speech tasks, but only 3 recording sessions, reviewed in [19]), or on a small number of speakers on multiple repetitions of isolated sequences (e.g. Whalen et al.
[20]: 1 speaker, 300 repetitions of monosyllabic words; Tielsen [21]: 6 speakers, 400 repetitions of the sequence /ipha/). We focused on the observation of intra-speaker variability in a restricted cohort of 9 speakers over a relatively large number of sessions, and on a read speech content allowing the characterization of productions on several dimensions using 21 acoustic descriptors. To make sense of this variability, it was compared to the variability that can be observed between two readings of different linguistic material or between speakers of the cohort.

The similarity observed between the descriptors showing the most variation between sessions and those varying the most between speakers raises serious questions about the influence of the recording session chosen in speaker separation tasks, even when the time-lapse between consecutive sessions is limited to days or weeks.

Another avenue for future research is related to the behavior of descriptors capturing articulatory dynamics in frequency bands (sd(energy) measures and their modulation). Their stability from one session to the next and within the same reading suggests the interest of further investigating these characteristics of the “way” of articulating and co-articulating which seems to be specific to the speaker. Our next studies on this corpus will therefore investigate the variability of temporal coordination measures between sounds produced successively through measures of coarticulation and undershoot of acoustic targets, in read speech but also on the semi-spontaneous recordings that we have also recorded for the same speakers.

5. Acknowledgements

This work has been supported by the "Investissements d'Avenir” program ANR-10-LABX-0083 (Labex EFL) and the VoxCrim project (ANR-17-CE39-0016).

Figure 2: Distribution of features values for each session (1 to 10) and each of the 9 speakers. Only a selection of 9 features commented in the text are presented in this figure.
6. References


