Forensic Science International 238 (2014) 59–67

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison

Adrian Leemann *, Marie-José Kolly, Volker Dellwo

Phonetics Laboratory, Department of Comparative Linguistics, University of Zurich, Plattenstrasse 54, 8032 Zürich, Switzerland

ARTICLE INFO

Article history:
Received 5 July 2013
Received in revised form 10 February 2014
Accepted 18 February 2014
Available online 5 March 2014

Keywords:
Speaker-individuality
Prosody
Suprasegmental temporal features
Speaking style variability
Channel variability.

ABSTRACT

Everyday experience tells us that it is often possible to identify a familiar speaker solely by his/her voice. Such observations reveal that speakers carry individual features in their voices. The present study examines how suprasegmental temporal features contribute to speaker-individuality. Based on data of a homogeneous group of Zurich German speakers, we conducted an experiment that included speaking style variability (spontaneous vs. read speech) and channel variability (high-quality vs. mobile phone-transmitted speech), both of which are characteristic of forensic casework. Speakers demonstrated high between-speaker variability in both read and spontaneous speech, and low within-speaker variability across the two speaking styles. Results further revealed that distortions of the type introduced by mobile telephony had little effect on suprasegmental temporal characteristics. Given this evidence of speaker-individuality, we discuss suprasegmental temporal features' potential for forensic voice comparison.

1. Introduction

When participating in a conversation with a group of people, we often have no trouble segregating speakers by their voices even if we have never met these speakers before. Experience also shows that we can typically identify an acquaintance on the telephone after only a few syllables. These examples illustrate that human voices are highly individual. The phenomenon that the speech signal contains speaker-individual information is exploited in speaker identification and verification procedures [1] and in particular in forensic voice comparison (hereafter FVC; [2]). In typical FVC cases, acoustic trace material from a crime (normally recordings of a perpetrator) are compared to acoustic comparison material (typically recordings of a suspect), and used for post-crime forensic investigations.

Voices can be individual in different acoustic domains. Research on speaker-individuality tended to focus on the frequency domain (fundamental frequency: [3–7]; formant frequencies: [8–16]) and the intensity domain of speech [17]. Relatively little attention has been paid to speaker-specific temporal characteristics. The principal objective of the present study is to examine in further detail the speaker-individuality of temporal features. Here we focus in particular on suprasegmental temporal features, that means temporal features of speech that are not restricted to a single segment (for example a consonant or a vowel, cf. [18]) but to the more global temporal organization of speech in an utterance. Such temporal organization of speech has traditionally been referred to as *speech rhythm*. For this reason, the aim of the present approach was to use measures that are frequently used in the field of speech rhythm.

Why do we focus on suprasegmental temporal characteristics?

(1) Effects of between-speaker suprasegmental temporal variability were observed for a number of different datasets in the field of speech rhythm [19–23]. These results, however, are based on datasets that were designed to investigate between-language effects [23] or they are based on a small number of speakers and on speech material in which possible between-speaker artifacts such as accent or dialect were not carefully controlled for [19–22]. In the present study, we investigated speaker-individual suprasegmental temporal features for 16 speakers that were highly controlled for accent (Zurich German), age (20–30 years), and social background (university students).
(2) Recent evidence points to two possible sources for speaker-individuality in suprasegmental temporal features: speaker idiolect and speaker anatomy. Speakers vary in an acquired way they use speech, having their own way of lengthening

* Corresponding author. Tel.: +41 44 634 59 48; fax: +41 44 634 43 57.
E-mail addresses: adrian.leemann@pholab.uzh.ch (A. Leemann), marie-jose.kolly@pholab.uzh.ch (M.-J. Kolly), volker.dellwo@uzh.ch (V. Dellwo).

http://dx.doi.org/10.1016/j.forsciint.2014.02.019
0379-0738/

sound patterns or having a preference towards certain syllables or sound segments. In terms of speaker anatomy, albeit coming from different strains of research, research showed that human movement is highly individual ([24,25] for gait; [26,27] for typing-movements). Eriksson and Wretling [28] suggested a comparable stability of timing patterns in human speech, which is in the same way produced by intricate, brain-controlled muscle movements as leg or finger movements found in walking and typing. It thus seems plausible that similar individualities as in human gait or finger movements are also present in articulatory movements.

In typical forensic phonetic casework the phonetic expert compares a number of speaker-individual characteristics between acoustic trace and comparison material. Such characteristics can either be on born features of the vocal tract such as voice fundamental frequency and vocal tract resonance characteristics or acquired features such as accent, dialect or sociolectal ways of pronunciation. It is essential for FVC to accumulate as many characteristics of the speech signal as possible [1,2]. With our research we aim at examining further speaker-individual information that may be used in FVC in the future. For such an application it is essential to have in-depth knowledge about the variability of the characteristics under scrutiny within and between speakers and about how such variables are affected by different signal transmission conditions (e.g. mobile phone). It is desirable for forensic circumstances that the features under investigation reveal maximum between-speaker variability and minimum within-speaker variability [1].

In forensic phonetic casework, trace and suspect material is either spontaneously produced or read. At the German Federal Criminal Police Office (BKA), an estimated 10–20% of trace material is read, while a somewhat larger amount, 10–30%, of suspect material is read (Olaf Köster, BKA, personal communication). The vast majority in both trace and suspect material, however, is spontaneously produced. Spontaneous and read speech differ on various levels: the former is optimized for human-to-human communication, shows simultaneous planning and execution, and overall demonstrates greater segmental and suprasegmental variability [19,29–31]. Since trace and suspect material often differ in speaking style, it is critical to know whether such differences affect the speech parameter being evaluated in FVC.

Aside from possible differences in speaking styles, trace and suspect material may also differ in channel transmission. 90% of the time, forensic trace and suspect material involves telephone-transmitted speech [32]. Telephone-transmitted speech is different from high-quality recorded speech in that it features band-pass transmission channels of only 350–3400 Hz [33,34], higher F1s, more narrow dynamic ranges, and artefactual peaks [33–36]. Mobile phone-transmitted speech shows wider variability in the transmission quality and more restrictive band pass filters, causing F1 in close and mid vowels to be even higher than over landline-telephone-transmitted speech [33]. Such technical effects compromise the reliability of frequency-based measures. It is likely that suprasegmental temporal features are advantageous in this respect: the points in the speech signal where vowels or consonants start, or where voicing starts or ends, should largely remain unaffected by the technical effects of mobile phone transmission. In this regard, suprasegmental temporal measures may be able to enhance FVC analyses particularly when the speech signal is degraded.

In a within-subject design, we examined the above-mentioned types of variability pertinent to forensic phonetics: speaking style variability (spontaneous vs. read) and channel variability (high-quality vs. mobile phone-transmitted). The design contained the same linguistic material, i.e. sentences, for each speaker and condition since temporal characteristics are known to be sensitive to sentence material [20,23]. The present study addresses the following research questions:

1. Are there between-speaker differences in suprasegmental temporal features? (see Section 3.1).
2. Which suprasegmental temporal measure explains most variation between speakers? (see Section 3.2).
3. How robust are suprasegmental temporal features to speaking style variability? (see Section 3.3).
4. How robust are suprasegmental temporal features to channel variability? (see Section 3.4).

Given the discussion above we expect to find significant between-speaker variability in both read and spontaneous speech as well as little within-speaker variability across the two speaking styles. Moreover, channel variability is likely to have little effect on suprasegmental temporal features.

2. Methods

2.1. Speakers

16 speakers (8 male/8 female) of Zurich Swiss German were recorded in a sound-treated booth at the Phonetics Laboratory of the University of Zurich. Eligibility criteria required individuals to demonstrate little to no regional and social accent variability. Average age was 27, SD = 3.6, and age range 20–33. None of the speakers reported hearing or speech disorders. The data was recorded in a sound-treated booth using a Neumann STH-100 transducer microphone (sampling rate of 44.1 kHz; 16 bit quantization).

2.2. Material

2.2.1. High-quality spontaneous speech

In a first recording session, spontaneous speech material was collected via semi-structured interviews. The 16 speakers were asked to talk freely to the interviewer (first and second author) about their studies at the University of Zurich. The interview was conducted in Swiss German. A subset of sentences, 16 per speaker (typically 15–20 syllables per sentence), was isolated from these interviews. For the isolation of sentences there are no formal criteria that allow for an identification of utterances as complete "units". We selected sentences according to syntactic, prosodic, voice quality, pausing and breathing criteria [37]. The isolated sentences had to form meaningful units and be fluently spoken, i.e. free from filled and unfilled pauses, hesitations, and mispronunciations. These 256 sentences (16 speakers × 16 sentences) formed the spontaneous, high-quality (henceforth *hifi*), corpus of this study.

2.2.2. High-quality read speech

We made orthographic transcripts of these 256 spontaneous sentences. These transcripts were given to the same 16 speakers with the request to prepare reading the sentences for a second recording session. Approximately three months after the first session, the 16 speakers read those 256 sentences in our laboratory (16 previously self-produced sentences + 240 sentences from their peers). It is plausible to assume that, given the temporal discrepancy between the two recording sessions, the obtained effects may either be attributed to speaking style variability or to the temporal delay between the two sessions. There is evidence from previous research, however, showing no effect of *test–retest* for suprasegmental temporal features [38]. These 4096 sentences (256 sentences × 16 speakers) constituted the read, hifi corpus of this study. The spontaneous and read hifi corpora amount to 56,794 syllables in total.