Speech Communication 86 (2017) 24–41

Contents lists available at ScienceDirect

Speech Communication

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

The beauty in a beast: Minimising the effects of diverse recording quality on vowel formant measurements in sociophonetic real-time studies

Tamara Rathcke[a,*], Jane Stuart-Smith[b], Bernard Torsney[c], Jonathan Harrington[d]

[a] English Language and Linguistics, University of Kent, UK
[b] English Language, University of Glasgow, UK
[c] School of Mathematics and Statistics, University of Glasgow, UK
[d] Institute of Phonetics and Speech Processing, University of Munich, Germany

ARTICLE INFO

Article history:
Received 12 May 2016
Revised 31 October 2016
Accepted 4 November 2016
Available online 9 November 2016

Keywords:
Real-time corpus
Formants
Formant normalisation
Noise
SNR
Spectral tilt
Sociophonetic ‘gold standard’

ABSTRACT

Sociophonetic real-time studies of vowel variation and change rely on acoustic analyses of sound recordings made at different times, often using different equipment and data collection procedures. The circumstances of a recording are known to affect formant tracking and may therefore compromise the validity of conclusions about sound changes made on the basis of real-time data. In this paper, a traditional F1/F2-analysis using linear predictive coding (LPC) was applied to the vowels /i u a/ extracted from spontaneous speech corpora of Glaswegian vernacular, that were recorded in the 1970s and 2000s. We assessed the technical quality of each recording, concentrating on the average levels of noise and the properties of spectral balance, and showed that the corpus comprised of mixed quality data. A series of acoustic vowel analyses subsequently unveiled that formant measurements using LPC were sensitive to the technical specification of a recording, with variable magnitudes of the effects for vowels of different qualities. We evaluated the performance of three commonly used formant normalisation procedures (Lobanov, Nearey and Watt-Fabricius) as well as normalisations by a distance ratio metric and statistical estimation, and compared these results to raw Bark-scaled formant data, showing that some of the approaches could ameliorate the impact of technical issues better than the others. We discuss the implications of these results for sociophonetic research that aims to minimise extraneous influences on recorded speech data while unveiling gradual, potentially small-scale sound changes across decades.

1. Introduction

1.1. On the issue of comparability in sociolinguistic data

Since its origins in the early 1960s, variationist sociolinguistics has been concerned with the methodological rigour of its quantitative enquiry. In the foreground of the early discussions were the issues primarily involving the data collection, such as the “Observer’s Paradox”, style shifts and sampling strategies (Labov, 1972; Cukor-Avila, 2000). Subsequent studies have further unveiled the multitude of the potential sources of influences in sociolinguistic data, which include (and are not limited to) familiarity between the participant and the interviewer, presence of additional peers during the interview, the experience and elicitation strategies of the interviewer as well as the quantitative approaches to analysing the data (Gregersen and Barner-Rasmussen, 2011; Labov, 1972; Milroy, 1987; Milroy and Gordon, 2008; Llamas et al, 2006; Tagliamonte, 2006; see Tillery and Bailey (2003) for a critical overview). All of these factors may not only influence the observed results, thus misleading generalisations about the patterns of variation and change, but also reduce comparability of the results across different studies of the same sociolinguistic phenomena, undermining the core principles of methodologically sound research, reliability and intersubjectivity.

Ultimately, sociolinguistic research aims to combine natural (or at least naturalistic) data which preserves the social identity (Scobbie and Stuart-Smith, 2012) with a rigorous amelioration of any extraneous influences that can influence the data patterns. In their critical paper, Tillery and Bailey (2003) suggested that this standard can only be achieved through a solid understanding of the sources and the magnitudes of possible extraneous influences on sociolinguistic data patterns, and regretted the current lack of such understanding, calling for more research in this methodologically highly relevant area.

[*] Corresponding author.
E-mail address: T.V.Rathcke@kent.ac.uk (T. Rathcke).

http://dx.doi.org/10.1016/j.specom.2016.11.001
0167-6393/Crown Copyright © 2016 Published by Elsevier B.V. All rights reserved.

The present study aims to contribute to this endeavour, and is concerned with the potential influences of technical specifications of recordings on the vowel formant measurements taken from them. Vowel formants are the core acoustic correlates of vowel quality typically obtained in sociophonetics (but see Harrington et al. (2013) for an alternative set of acoustic measures), and have been scrutinised in many studies of sound variation and change (e.g. Fought, 1999; Gregersen et al., 2009; Harrington et al., 1997; Labov, 1994; Labov et al., 2006; Maclagan et al. 2009; Mesthrie 2010). In an apparent-time setting, much care has traditionally been taken to account for the formant differences arising from speaker physiology, relating primarily to the age and the vocal tract size (e.g. Linvillea and Rens, 2001), and to distinguish these physiological influences from the sociolinguistically relevant patterns produced by speakers of different ages and sexes (e.g. Labov et al., 2006). Numerous techniques have been developed, tested and compared in order to achieve the normalisation for speaker physiology while preserving the social indexicality of their speech (e.g. Adank et al., 2004; Clopper, 2009; Watt and Fabricius, 2002; see Flynn (2011) for an overview). We will discuss the most commonly used approaches in Section 3.3 below.

In contrast to this long-standing methodological debate characteristic of apparent-time studies, real-time studies of sound variation and change have rarely problematized potential issues involved in formant measurements of vowels. Trend studies with real-time data (recorded with different samples of individuals from the same community at different points in time) are unanimously recognised as a particularly insightful and reliable methodological setting for studying language change at a community level (e.g. Labov, 1994; Sankoff and Blondeau, 2007; Trudgill, 1988), primarily because they eliminate effects related to speaker age, such as age grading (Wagner, 2012). However, real-time studies frequently rely on acoustic analyses of recordings of speech made using different equipment with variable technical specifications and following different recording procedures. To date, still little is known about the sources, types and magnitudes of technical influences on the formant data. In the next section, we will give an overview of the currently established effects, and hypothesise how they might play out in a real-time study of sound variation and change.

1.2. Technical influences on formant measurements

Not many studies have addressed the question of whether, and how, formant values (extracted using the traditional method of LPC) might be influenced by the equipment and set-up of a recording and its resulting technical specifications. A series of studies have been conducted in the context of forensic speaker identification (e.g. Byrne and Foulkes, 2004; Künzel, 2001); and only a few, mostly preliminary investigations have recently pointed out that technical issues of a recording may obscure the patterns of variation and change in sociophonetics, too (De Decker and Nycz, 2011; De Decker, 2016; Hansen and Pharao, 2006; Hansen and Pharao, in progress).

In terms of the recording equipment and set-up, several features have been identified to leave an imprint in the vowel spectrum and to impact on the measured formant values. First of all, the band-pass filtering due to the transmission by phone lines (both mobile and landline) is known to interfere with the calculation of the formants (Byrne and Foulkes, 2004; Künzel, 2001). Harmonics that lie below the lower cut-off boundary (approximately 300 Hz) and above the upper boundary (approximately 3.2 kHz in mobile phones and 3.5 kHz in landline transmissions) are most affected, since their weighting in the calculation of the formant frequencies is decreased. This usually leads to artificially high frequencies of F1 (particularly in high vowels whose F1 is much stronger affected than the relatively high F1 of low vowels). However, even F2 whose frequencies fall within the transmitted range shows some technically introduced artefacts. In comparison to the values obtained from a recording made simultaneously with a studio microphone, F2 of high vowels tends to measure lower values in mobile recordings (Byrne and Foulkes, 2004), though the effect tends to be smaller and has not been consistently documented in other phone transmissions (Künzel, 2001). The exact magnitudes of these technically introduced effects also seem to vary substantially across different studies and types of phone transmissions, and range between 14 and as high as 60 percent of the original frequency (Byrne and Foulkes, 2004; Künzel, 2001).

Similar to the effects of band-pass filtering for a cost-effective phone transmission, compression algorithms used for a space-effective storage of video and digital audio recordings (as e.g. available on the internet) have been shown to influence spectral properties of speech recordings (De Decker and Nycz, 2011; Rozborski, 2007; van Son, 2005). F1 seems to be affected across the board, measuring higher values after a compression, while the impact on F2 is rather mediated by vowel quality, raising F2 in high vowels but lowering it in low vowels (De Decker and Nycz, 2011). Again, the magnitude of these effects varies across studies and compression methods, ranging from negligible (≤3%, van Son, 2005) to quite substantial (De Decker and Nycz, 2011), with higher compression rates leading to a more significant distortion of the original recording (Rozborski, 2007). Although mobile devices admittedly introduce numerical artefacts in the formant values during the transmission (cf. Byrne and Foulkes, 2004), De Decker and Nycz (2011:54) argue that recordings made with some portable devices of the same manufacturer (here, Macbook Pro and iPhone) produce comparable measurements, and maintain an overall shape and size of the vowel space in comparison to uncompressed recordings (at least as far as F1 and F2 are concerned), thus lending themselves to a sociolinguistic investigation better than others (e.g. Mino-derived formats commonly used by YouTube).

Apart from the influence the format of a recording can have on its spectra and formant measurements taken using LPC, somewhat less obvious factors, such as ambient noise, room acoustics, microphone make and placement during the recording session, have also been shown to leave their spectral imprints and interfere with formant measurements (De Decker, 2016; Hansen and Pharao, 2006, Hansen and Pharao, in progress; Plichta, 2004; Svec and Granqvist, 2010). The quality of the recordings not controlled for such influencing factors will likely vary with respect to at least two technical specifications (cf. Svec and Granqvist, 2010): (1) the levels of noise, typically measured by the signal-to-noise ratio, SNR (see 2.3) and (2) spectral balance (or tilt), reflected in the distribution of the intensity across lower vs. higher harmonics of the spectrum (see 2.3 for more detail).

It is well known that high levels of background noise reduce intelligibility of speech (e.g. Pollack and Pickett, 1958), but even recordings made in relatively quiet surroundings can differ with respect to their SNR. For example, hiss (or low-level white noise) can originate from analogue electronics, ground hum and buzz from improperly grounded systems: the fundamental of 50 or 60 Hz and their harmonics will be distinguishable in the recording spectrum (Corley, 2010). An increased distance of the microphone from the sound sources can also decrease SNR, making the room reverberation and noises more prominent in a sound recording (Corley, 2010:57). Omnidirectional microphones usually pick up more background noise than directional ones, with the small-tip versions producing particularly noisy recordings (Svec and Granqvist, 2010). In such increased noise levels (reflected in lower SNR, see 2.3), formants often appear very faint or have larger bandwidths and are therefore less clearly defined (Plichta, 2004); Plichta strongly advises against using such recordings for speech research. De Decker (2016), however, shows that not all types of background noise have