INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Filling the population statistics gap: Swiss German reference data on F0 and speech tempo for forensic contexts

Hannah Hedegard[1], Andrea Fröhlich[2], Fabian Tomaschek[3], Carina Steiner[1], Adrian Leemann[1]

[1]Universität Bern
[2]Zurich Forensic Science Institute
[3]Universität Tübingen

hannah.hedegard@unibe.ch, Andrea.Froehlich@for-zh.ch, fabian.tomaschek@uni-tuebingen.de, carina.steiner@unibe.ch, adrian.leemann@unibe.ch

Abstract

The increased focus on big data in phonetics, and Bayesian statistics in the forensic sciences, prompt a fundamental issue in common applications of forensic phonetics. Relevant population distributions for most features, a key element when evaluating the similarity and distinctiveness of voices, remain lacking for a substantial number of languages and dialects. This paper provides population statistics for two phonetic features in the Swiss German context, speech tempo and F0, and outlines a potential method for big data analysis. The speech data is taken from 1000 SwG speakers and include two different style conditions: spontaneous and read speech. Results indicate significant variation for both parameters: we contradict previous findings on gender differences in speech tempo and note discrepancies for both features between the two styles. These findings constitute an important contribution to the field of forensic phonetics, as well as the field of general phonetics more broadly.

Index Terms: Forensic Phonetics, population statistics, Swiss German

1. Introduction

The advent of big data in the field of phonetics has led to it being described as both the solution and the problem: large datasets produce more reliable results but necessitate robust methods that can handle quantitative analysis on such a scale. Phoneticians remain behind with comprehensive big data-friendly digital techniques. For forensic phoneticians, though, the methodological gap prompted by an onus on greater sampling is particularly problematic due to the recognition of Bayesian statistics as the only scientifically sound approach for forensic casework. Recent work (e.g., [1]) has highlighted that when using Bayesian likelihood-ratios in forensic sciences, dependable and relevant reference data is of crucial importance. Hereafter we provide an example from forensic phonetic casework that demonstrates where one requires reference data:

A crime (such as a threat, fraud etc.) was recorded on a portable recording device and police investigations have led to a suspect speaker. A forensic phonetician now performs a Speaker Comparison analysis, including auditory-phonetic, acoustic and, in specific cases, automatic methods. The auditory-phonetic approach follows a detailed protocol that analyzes various features of a speaker's voice, language use and speaking style. When evaluating fundamental frequency (F0), the questioned recording and the comparison recording show a mean frequency of around 250 Hz and 260 Hz respectively. Besides F0 means, many F0-related features, such as min, max, standard deviation, range or variation coefficient are usually also assessed in a forensic speaker comparison case. The forensic caseworker needs to weigh up the value of the individual findings but is faced with several key questions:

- Do these values lie within one speaker's possible F0 variability range?
- Is the respective feature stable across speaker modalities and different channel conditions?
- What's the typicality of this feature across the relevant population?
- How do F0 values vary across different dialects, accents, ages and speakers?

To answer these questions, caseworkers need information on population statistics from reference data. Such data would preferably be comparable to the case in terms of speaking styles and recording conditions. If we can establish a reference distribution of a particular feature from this data, we have a clearer idea of how exceptional our obtained case value (e.g., 250 Hz) is, and thereby estimate how coincidental it is that the suspect and the criminal's voices have similar fundamental frequency values.

However, the key problem is the shortage of detailed and inclusive population statistics for phonetic parameters in most languages besides the larger, well-researched ones, and even in the case of the latter, often utilized methods do not allow for forensically applicable results. While there has been considerable analysis of phonetic features in e.g., English, Arabic, and Mandarin, our knowledge of how these parameters compare and vary in minority languages and dialects, such Swiss German (SwG), is relatively small. Previous investigations into the distribution of acoustic features in SwG have either relied on smaller datasets, both in terms of speaker numbers and regional/social scope, [2, 3, 4] or unnatural speech data [5]. To narrow this knowledge gap, we present results on speech tempo and fundamental frequency in SwG, yielded from a novel automated method of data extraction and processing. We base our findings on a large speech corpus, consisting of extended recordings with 1000 speakers from across German-speaking Switzerland, balanced for gender and age. We also compare two different recording conditions: spontaneous and read speech, which provides additional information on intra-speaker variability.

We will first describe the methods we used, then present the results and provide an extended discussion of our

10.21437/Interspeech.2023-1042

findings and their contribution to the fields of general phonetics and forensic speech science specifically.

2. Methods

2.1. Material

The material for the present study was obtained from the Swiss German Dialects Across Time and Space (SDATS) corpus [6]. The corpus contains recordings of 1000 speakers of SwG in two conditions. In the read condition, speakers read a text of 266 words. Speech in the spontaneous condition was taken from a 15-minute-long sociolinguistic interview. 85% of the speakers were recorded via their mobile phones, but we do not expect any confounding effect on F0 due to the range in mobile phone device type [7].

2.2. Speakers

Speakers for the corpus were recruited from all over German-speaking Switzerland. For the current analysis, a total of 987 speakers were analysed (13 speakers were excluded due to technical problems during pre-processing). Speakers were recorded from two age cohorts: 17 – 43 years and 53 – 95 years. For the present study, each age cohort was subdivided into two sub-cohorts by means of a median split, resulting in four age cohorts: young 1 (17 – 26 years, N(female) = 157, N(male) = 127), young 2 (27 – 43 years, N(female) = 9891, N(male) = 123), old 1 (53 – 69 years, N(female) = 134, N(male) = 117), old 2 (70 – 95 years, N(female) = 110, N(male) = 130).

2.3. Variables

The dependent variables for the present study – average speech tempo and average F0 values for each speaker – were yielded in two steps.

In the first step, speech tempo (ST) and F0 were extracted automatically. For ST, we used the PRAAT [8] script by de Jong et al. [9]. The script automatically detects intensity peaks and groups them into phrases separated by pauses longer than 300 ms. We used the default settings but changed the silence threshold to -20 dB, thereby excluding the interviewer's speech in the background. A manual check verified that the script is capable of locating syllabic nuclei with relatively high accuracy. Accordingly, a measure of ST was obtained by dividing the number of intensity peaks in a phrase by the phrase's duration (thus the measurement of peaks/sec). In the reading condition, the script found on average 52.6 phrases per speaker (sd = 17.4) with an average of 8.6 intensity peaks (sd = 2.4) per phrase and an average duration of 2.03 s (sd = 1.42 s). In the spontaneous condition, the script found on average 229.5 phrases per speaker (sd = 68.5) with an average of 7.4 intensity peaks (sd = 3.7) per phrase and an average duration of 1.67 s (sd = 1.69 ms).To obtain F0 measurements per speaker, we calculated F0 in Hertz in a window of +/- 1.5 ms around the time points of the intensity peaks using a custom made PRAAT script. A minimum F0 threshold of 50 Hz and maximum threshold of 600 Hz was kept constant across male and female speakers.

In the next step, we calculated the average ST on the basis of individual ST measurements and average F0 on the basis of individual F0 measurements for each speaker. To do so, we had to clean the data of interviewer phrases and those that were either too short to reliably provide F0/ST information or were erroneously marked by the script. We therefore only included in our analysis ST and F0 from phrases for which more than 5 peaks and an F0>0Hz were found. In the reading condition, the exclusion resulted in a data loss of 39.4% for ST, yielding a total of 31,898 SR measurements; and a loss of 14.3% for F0, yielding a total of 3,659,957 F0 measurements. In the spontaneous condition, the exclusion resulted in a data loss of 55.0% for ST, yielding a total of 103,161 ST measurements; and 18.6% for F0, yielding a total of 1,317,027 F0 measurements. Having discussed our methods, we will report the results of linear regression models testing these effects in the next section.

3. Results

3.1. Overview

Table 1: *Mean values for F0 and ST amongst the men*

Age Group	F0		Speech tempo	
	Spon.	Read	Spon.	Read
17-26	122.5	118.9	4.3	4.2
27-43	122.4	120.8	4.4	4.2
53-69	131.3	128.3	4.3	4
70-95	138.1	130.4	4.3	4

Table 2: *Mean values for F0 and ST amongst the women*

Age Group	F0		Speech tempo	
	Spon.	Read	Spon.	Read
17-26	203.5	209.6	4.4	4.3
27-43	201.8	202.4	4.4	4.2
53-69	195.7	196.2	4.4	4
70-95	191.5	192.5	4.4	3.9

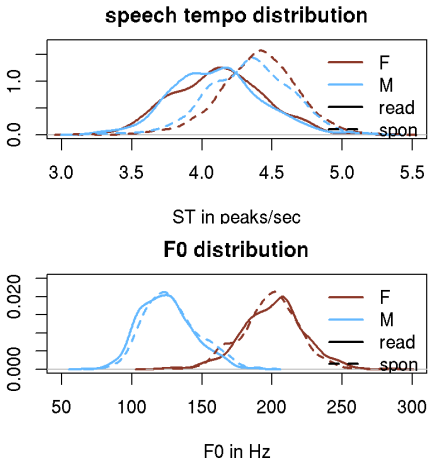

Figure 1: *Speech tempo (ST) and F0 distributions*

3.2. Speech tempo

As can be seen in Table 1 and Figure 1, we find that speakers in the spontaneous recording condition have a significantly higher SR than in the reading condition ($t = 18.7$, $df = 1954$, $p < 0.001$). Given that we have only one measurement per speaker, and given that the results are normally distributed,

all statistical results are based on t-tests, whenever only t-values are reported. Regarding differences between male and female, we do not find any significant difference in the reading condition, i.e. male and female speakers have the same speech tempo (t = 0.7, df = 986, p = 0.493). By contrast, in the spontaneous condition, women have a significantly higher speech tempo than men (t = 3.6, df = 982.9, p = 0.0003).

Next, we turn our attention to the effect of age on average speech tempo. In the following figures, we illustrate the age effect by means of the cohorts described in Section 2.2 but will use numeric age to predict prosodic characteristics in a linear regression model. In Figure 2, we observe that as female speakers become older, their SR becomes slower. The effect is stronger in the reading condition (β = -0.0071, se = 0.0005, t = -13.08, p < 0.0001) than in the spontaneous condition (β = -0.0013, se = 0.0005, t = - 2.345, p = 0.0189).

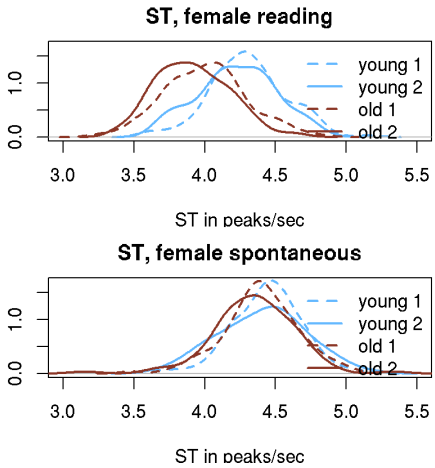

Figure 2: *Speech tempo distributions for female speakers depending on age and recording condition*

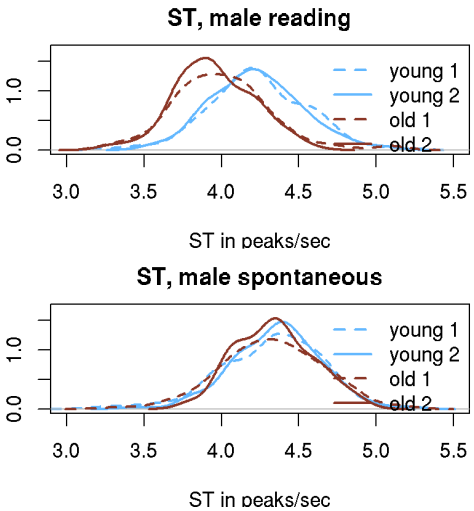

Figure 3: *Speech tempo distributions for male speakers depending on age and recording condition*

By contrast, in Figure 3 we observe that lower speech tempo in relation to higher age in male speakers is present only in the reading condition (β = -0.006, se = 0.0005, t = -10.31, p < 0.0001), but not in the spontaneous condition (β = -0.00027, se = 0.0006, t = -0.457, p = 0.648). When speakers are separated into two age cohorts (young 1: 17-26, and young 2: 27-43 vs. old 1: 53-69, and old 2: 70-95) in the reading condition, we find that the effect is not present in the young cohort (β =-0.0048, se = 0.0038, t = -1.244, p = 0.215) but only in the old cohort (β =-0.0092, se = 0.0029, t = -3.208, p = 0.00151).

3.3. F0

In regard to F0, we find (unsurprisingly) significant differences between male and female speakers in both the reading and the spontaneous condition (t = 58.3/54.3, df = 962/980, p < 0.0001/<0.0001), see Table 1 and Figure 1. We observe a strong overlap between the reading and spontaneous conditions for both male and female speakers which might indicate that there are no significant differences in F0 between the two recording conditions. Indeed, a Student's t-test finds a minimal difference for male speakers (124.6 Hz in reading vs. 128.6 Hz in spontaneous speech, t = -3.33, df = 978.93, p < 0.001), but not for female speakers (200.8 Hz in reading vs. 198.4 Hz in spontaneous speech, t = 1.78, df = 978.93, p = 0.077).

Moving on to the effect of age (Figure 4), we observe that older female speakers show lower F0. The effect is greater in the reading condition (β = -0.307, se = 0.042, t = -7.272, p < 0.0001) than in the spontaneous condition (β = -0.218, se = 0.041, t = -5.363, p < 0.0001). Conversely, male speakers (Figure 5) have higher F0 as they grow older. The effect is weaker in the reading condition (β = 0.224, se = 0.037, t = 6.119, p < 0.0001) than in the spontaneous condition (β = 0.3, se = 0.037, t = 8.127, p < 0.0001) .

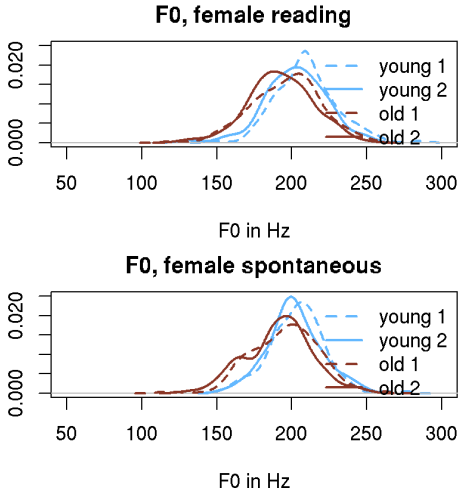

Figure 4: *F0 distributions for female speakers depending on age and recording condition*

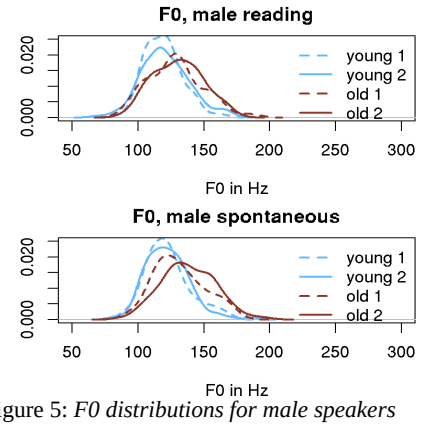

Figure 5: *F0 distributions for male speakers depending on age and recording condition*

4. Discussion

4.1. Speech tempo

The average speech tempos in this study (e.g., 4.4 syll/sec for a man speaking spontaneously) appear noticeably slower than have been found previously for SwG dialects such as those spoken in Bern (5.0 syll/sec) and Zürich (5.8 syll/sec) [5]. Articulation rates in Standard German have also been found to be considerably faster, with Jessen reporting values of 5.21 syll/sec in the read condition and 5.41 syll/sec in the spontaneous speech [10]. While these results may indicate there is some truth to the cliché that the Swiss speak slower than Germans, potentially the discrepancies are due to methodological differences, specifically that the speech data here is natural rather than single word utterances and calculated with different methods, respectively. For example, the threshold selection for intensity peak detection may affect speech tempo results.

Künzel found that German speakers from Germany were consistent across read and spontaneous conditions, suggesting articulation rate is a feature with "intra-individual stability" [11], while Jessen's read data was faster than the spontaneous [10]. Our data aligns with other studies of English that recorded a style effect of faster speech when it is natural rather than read, and we echo explanations that this is attributable to the careful attention and mental load involved in reading out loud [12]. In the SwG context, this effect is likely to be more stark given the sociolinguistic conditions of the language; as a primarily oral language, written SwG is rarely read out loud and is an unnatural task for many speakers, particularly amongst the older generation.

When we look at spontaneous speech alone, we see a clear gender discrepancy, with women speaking significantly faster than men, a pattern that contradicts previous findings in SwG [5] and other languages e.g., American English [13] and Standard German [14]. Ageing slows down women's speech, though more in the read speech style than spontaneous, while only read speech is slowed by age for men. This latter finding that the spontaneous speech of men is unaffected by age is unexpected given what other scholars have found for articulation rates in English [e.g., 15].

4.2. Fundamental frequency

In terms of interaction between social and biological factors we find both expected and unexpected results for fundamental frequency. Our data shows F0 was lower amongst older women, but higher for older men, a finding supported by studies of other languages [e.g., 16], and explained by hormonal changes across the lifespan. Interestingly, it seems as though male SwG speakers modulate their fundamental frequency according to speech style: men drop their fundamental frequency in reading in contrast to spontaneous speech. We posit that societal norms regarding femininity and masculinity are partly behind this pattern [17], though it is questionable whether this minimal difference is actually perceivable.

The effect of age provides conflicting support for this claim, as the effect of a higher F0 as men age is tempered in the reading condition (suggesting they are consciously lowering their voices in this style), but for older women, reading out loud in fact is more likely to result in a deeper voice.

4.3. Implications for Forensic Phonetics

Perfectly matching population statistics for each and every case is not a realistic goal. Casework will therefore require a combination of the experts' experience and - where they exist - statistical data. Knowing more about the distribution and robustness of features across different channels, age levels, dialects etc. will help to address the question of accurate statistical data. The data presented here on speech tempo and F0 therefore forms an important contribution for SwG forensic phonetic casework. Speech tempo and F0 are key phonetic parameters for analysis, and to date, no other studies have set out this information for both genders, across age groups and styles, using natural speech data – closely matching forensic phonetic conditions. Further work on forensically pertinent metrics of these features, e.g., F0 max, min, range, variation coefficient etc, would build on the foundations laid here.

5. Conclusion

Aside from providing valuable reference points for speech tempo and fundamental frequency in SwG, our findings suggest there are several notable speech tempo and F0 phenomena occurring in SwG that correspond to style, age and gender categories, and also confirm that some phonetic characteristics are shared with other languages i.e., are potential "universals".

In forensic phonetics, ideal conditions (recording, data, population statistics etc.) are extremely rare. Practitioners usually lack the time and resources to gather big databases alongside doing case work. We strongly advocate for practitioners and researchers to work closely together to fill the current gap of reference values for almost every minority language variety, a critical component of quantitative methods in forensic phonetic casework.

6. Acknowledgements

The authors thank the Zurich Forensic Science Institute for their collaboration, and the SNF (PCEFP1_181090).

7. References

[1] G. S. Morrison, "Advancing a paradigm shift in evaluation of forensic evidence: The rise of forensic data science," *Forensic Science International: Synergy*, vol. 5, Article 100270, May 2022.

[2] J. Fleischer, S. Schmid, "Zurich German," *Journal of the International Phonetic Association*, vol. *36*, issue 2, pp. 243–253, Dec. 2006.

[3] A. Leemann, B. Siebenhaar, "Intonational and temporal features of Swiss German" in Proc. *16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, Aug. 2007, pp. 957-960.

[4] A. Leemann, V. Dellwo, M. Kolly, S. Schmid, "Rhythmic variability in Swiss German dialects," in Proc. *6th International Conference on Speech Prosody*, Shanghai, May 2012, pp. 607-610.

[5] A. Leemann, "Analyzing geospatial variation in articulation rate using crowdsourced speech data," *Journal of Linguistic Geography*, vol. 4, issue 2, pp. 76-96, Sep. 2016.

[6] A. Leemann, P. Jeszenszky, C. Steiner, M. Studerus, & Messerli, J. "Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing," *Linguistics Vanguard*, vol. 6, no. 3, pp. 1-16, Sep. 2020.

[7] S. Jannetts, F. Schaeffler, J. Beck, S. Cowen, "Assessing voice health using smartphones: bias and random error of acoustic voice parameters captured by different smartphone types," *International Journal of Language and Communication Disorders*, vol. 54, issue 2, pp. 292-305, Feb. 2019.

[8] Praat: doing phonetics by computer. (Version 5.1.05), P. Boersma, D. Weenink. Accessed: May 1, 2009.

[9] N. H. De Jong, J. Pacilly, W. Heeren, "PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically," *Assessment in Education: Principles, Policy & Practice*, vol. 28, issue 4, pp. 456-476, Jun. 2021.

[10] M. Jessen, "Forensic reference data on articulation rate in German," *Science & justice: journal of the Forensic Science Society*, vol. 47, issue 2, pp. 50–67, Sep. 2007.

[11] H. J. Künzel, "Some general phonetic and forensic aspects of speaking tempo." *International Journal of Speech, Language and the Law*, vol. 4, issue 1, pp. 48–83, Jul. 1997.

[12] E. Jacewicz, R. A. Fox, L. Wei, "Between-speaker and within-speaker variation in speech tempo of American English," *Journal of the Acoustical Society of America*, vol. 128, issue 2, pp. 839-850, Aug. 2010.

[13] T. H. Crystal, A. S. House, "Articulation rate and the duration of syllables and stress groups in connected speech," *Journal of the Acoustical Society of America*, vol. 88, issue 1, pp. 101-112, Jul. 1990.

[14] A. P. Simpson, "Phonetische Datenbanken des Deutschen in der empirischen Sprachforschung und der phonologischen Theoriebildung," in Arbeitsberichte AIPUK 33, Kiel, Nov. 1998.

[15] E. Jacewicz, R. A. Fox, C. O'Neill, J. Salmons, "Articulation rate across dialect, age, and gender," *Language Variation and Change*, vol. 21, issue 2, pp. 233-256, Jul. 2009.

[16] E. T. Stathopoulos, J. E. Huber, J. E. Sussman, "Changes in Acoustic Characteristics of the Voice Across the Life Span: Measures From Individuals 4-93 Years of Age," *Journal of speech, language, and hearing research*, vol. 54, issue 4, pp. 1011-1021, Aug. 2011.

[17] C. T. Ferrand, R. L. Bloom, "Gender differences in children's intonational patterns," *Journal of Voice*, vol. 10, issue 3, pp. 284-291, May 1996.