“Who’s Calling, Please?” Speaker-Specific Information in Vowels

Pieternel van Braak (3924076)

BA Scriptie Taalwetenschap (TW3V14002)

Supervisor: dr. W.F.L. Heeren

Second reader: dr. H. Quené

10 April 2015

Universiteit Utrecht

Abstract

The present study investigates the effectiveness of long-term formant distributions (LTFDs) and by-vowel long-term formants (LTFs) as measures in speaker comparison, by means of a small-scale perception experiment and acoustic analysis of speech samples of four comparable speakers. Overall, LTFD3 is suggested to be a more effective speaker discriminant parameter than LTFD2 in both telephone recorded and directly recorded samples. Both LTFD2 and LTFD3 measurements effectively discriminated between all pairs of speakers, although the occurrence of within-speakers differences between distributions emphasizes the need for further research into the use of LTFD[2,3] as a measure in speaker comparison. Furthermore, the results suggest that analyses of individual vowel spaces in directly recorded samples could be a useful addition to the arsenal of speaker discriminant measures in speaker comparison studies.

Contents

1. Introduction 4

2. Method 7

2.1. Acoustic analysis 7

2.1.1. Participants 7

2.1.2. Materials 7

2.1.3. Procedure and data analysis 8

2.2. Perception experiment 8

2.1.1. Participants 8

2.1.2. Materials 9

2.1.3. Procedure and data analysis 10

3. Results and analysis 10

3.1. Influence of direct versus telephone recording on auditory speaker identification 10

3.2. Applicability and effectiveness of LTFD[2,3] 12

3.3. LTFD2 in sets of front and back vowels 14

4. Discussion and conclusion 15

5. Acknowledgements 15

6. References 19

7. Appendices 21

A. Orthographic transcription of recorded passage 21

B. Mean, SD and 95% CI of F2 per sample per speaker 21

C. Kolmogorov-Smirnov Z values for LTFD[2,3] per speaker pair 22

D. Vowel spaces in direct recording per speaker 23

1. Introduction

Acoustic analysis of speech samples in forensic speech science aims to find features of speech that allow for effective speaker comparison. Essentially, effective acoustic measures for speaker comparison should be characterized by maximum inter-speaker variability coupled with minimum intra-speaker variability (Gold, French, & Harrison, 2013). The present study focuses on acoustic analysis of vocalic content; it investigates the effectiveness of long-term formant distributions (LTFDs) and by-vowel long-term formants (LTFs) as measures in speaker comparison. LTFDs are defined as distributions of frequency values for each formant based on all vowels produced by a speaker in a speech sample. LTF(D)s in general are proposed as a useful measure in several speaker comparison studies (cf. Nolan & Grigoras, 2005; Moos, 2010; Gold et al., 2013).

In this study, LTF(D) measurements in both monozygotic twins' speech samples and non-twins' speech samples are expected to shed light on the effectiveness of LTF(D) in distinguishing between samples of maximally similar pairs of speakers. Monozygotic twin pairs are assumed to be characterized by maximum anatomical similarity and maximally equal environmental experience (San Segundo Fernández, 2014). Although the degree of differences is not always equal across twin pairs for all parameters (Loakes, 2006), generally speaking differences in acoustic information between twin pairs are presumably small. For this reason, twin speech is suited for testing the distinctive power of speaker comparison parameters.

An anecdotal situation suggests that telephone transmission of speech might influence perceptual qualities of the speech signal. The father of a monozygotic twin pair reports that, at times, he is unable to identify his adult twin daughters in telephone communication, even though correct identification in direct communication yields no problems for him (personal communication, 2014). Since this is an anecdotal situation, there is no evidence that differences in acoustic information cause crucial perceptual differences between transmission channels. However, the literature suggests that the filtering effect of telephone transmission indeed influence acoustic qualities of the speech signal. According to Künzel (2001) and Rose (2003), telephone transmission of speech particularly affects frequencies below 300 Hz and above 3,400 Hz. Considering these findings, LTF(D)[1,4,5] are not investigated in the present study; only LTF(D)2 and LTF(D)3 are investigated. Since higher vowel formants are assumed to encode relatively much speaker-specific information (cf. Jessen & Becker, 2010; Gold et al., 2013), it is expected that the filtering effect of telephone transmission especially influences the effectiveness of LTF(D)3 as a measure in speaker comparison.

Supposedly, anatomical and environmental similarities between monozygotic twins, in combination with the filtering effect of telephone transmission (cf. Byrne & Foulkes, 2004) may influence the speech signal in such a way that speaker identification is severely obstructed. The current research aims to explore to what extent transmission channel characteristics of direct versus telephone recording influence the success rate of auditory speaker identification in a monozygotic twin pair. This question is addressed by means of a small-scale perception experiment, in which a third (non-related) speaker is included to allow for comparison of speaker confusion between twin and non-twin speakers in both transmission channels.

Due to the assumed minimum inter-speaker variability between monozygotic twins and the generalizing quality of LTFs, it is expected that LTFD[2,3] may not be powerful enough to distinguish between speech samples of monozygotic twins. Several studies suggest that certain categories of vowels might contain more speaker-specific information than the vowel inventory as a whole (e.g. Loakes, 2004; Stevens et al., 1968; Dukiewicz, 1970; Pickett, 2003). For example, the Quantal Theory of speech (Stevens, 1989) implicates that the cardinal vowels /i, a, u/ might contain more speaker-specific information than other vowels, because intra-category variability in cardinal vowel categories is presumably lower than for other vowels (cf. Stevens, 1989) and because the cardinal vowels are assumed to be “in approximately the same location [in the vowel spaces] across all languages” (Al-Tamimi & Ferragne, 2005, p. 2465).[1] In a different approach to vowel-specific degrees of speaker-specificity, Loakes (2004) relates the effectiveness of LTFDs as a speaker comparison measure to the place of articulation of specific vowels. She states that “researchers have found that […] especially front vowels and close-front vowels in particular, are more useful than other parameters for highlighting speaker-specificity” (p. 289), because they have F2s in the higher spectral region. Expanding on the abovementioned line of research and considering the observed influence of vowel-

[1] Although both Bradlow (1995) and Al-Tamimi and Ferragne (2005) confirm the last statement, it is disputed by Engstrand and Kull (1991) in their comparative study including seven languages.

specific places of articulation (which is related to the second formant), the present study compares speaker-specific aspects of the position of a set of back vowels with the position of a set front vowels in the vowel space in direct recording.

In conclusion, the present paper expands on previous investigations into the effectiveness of LTF(D)[2,3] as a measure in speaker comparison in different contexts. The following research questions are posed:

Research question 1. To what extent do the transmission characteristics of *direct recording* versus *(landline) telephone recording* influence the success rate of auditory speaker identification in a monozygotic twin pair? Confusion is expected to be higher between twin speakers than between non-twin speakers, and higher in telephone speech relative to studio speech, since a) in comparison with direct recording, acoustic information in the lowest and higher spectral regions of the speech signal is lost due to the filtering effect of telephone transmission (Künzel, 2001; cf. Byrne & Foulkes, 2004); and b) anecdotal information suggests that auditory speaker identification in a monozygotic twin pair is more successful in direct communication than in telephone transmitted communication.

Research question 2. To what extent are LTFD[2-3] applicable and effective as parameters in speaker comparison? It is expected that a) LTFD-2 and LTFD-3 (cf. Byrne & Foulkes, 2004) will yield similar distances between speakers in direct recording compared to telephone recording; b) given that LTFD-3 seems to be most effective in speaker identification (Moos, 2010; Gold, French, & Harrison, 2013), LTFD-3 distances between speakers will be larger than LTFD-2 distances between speakers; and that c) for LTFD[2,3], distances within speakers will generally be smaller than distances between twins within a monozygotic twin pair, and that LTFD[2,3] distances between speakers of non-twin pairs of speakers will be largest of all.

Research question 3. To what extent does by-vowel analysis of vowel subsets and analysis of (relative) positions of individual vowels in the vowel space contribute to the use of across-vowel LTFD as a measure in speaker comparison? A separate analysis of (specifically) LTFD-2 in front vowels is expected to be an effective speaker discriminant measure in addition to LTFD[2,3] analysis across vowels. Also, it is expected that analysis of the distribution of individual vowels in the vowel space might reveal more speaker-specific details than analysis of LTFD[2,3] across vowels on its own.

2. Method

2.1. Acoustic analysis

2.1.1. Participants

Four female native speakers of Dutch (one identical twin pair, and two unrelated speakers), aged between 20 and 22 participated. The speakers in the investigation are MB and PB (identical twin pair), AK and FW (unrelated participants). All speakers are university students. The twins share the same education until the end of high school (to 18 years of age), and the non-twin speakers share comparable education both with the twin pair and with the other non-twin speaker. The speakers share intermediate to advanced L2 proficiency in English. All speakers were raised in the Dutch province Utrecht; except for speaker FW, who moved from Zuid-Holland to Utrecht when she was 4 years old. All speakers are non-smokers (cf. Gonzalez & Carpi, 2004).

2.1.2. Materials

The data consist of recordings of a reference text in Standard Dutch containing all vowels of the language. The reference text used is *De noordenwind en de zon* (as provided in Gussenhoven, 1992)[2]. The speakers were recorded in a quiet room at the UiL-OTS laboratory of Utrecht University. Telephone transmitted recordings were obtained via a telephone connection with the Netherlands Forensic Institute in The Hague. Telephone transmitted recording and direct recording took place simultaneously, and there was a short practice session with a different text. Each speaker was requested to read the text twice, resulting in four samples of read speech per individual (2 versions x 2 types of recording). All speech samples are at least 30 seconds in duration (it should be noted that this is longer than the recommended minimum of approximately 19 seconds of read speech for LTF analysis, as proposed in Moos, 2010).

Studio recordings were made on an Audio-Technica AT8410a microphone, positioned approximately 25cm from the participants' mouth. Telephone transmitted speech samples were recorded at the Netherlands Forensic Institute. Each telephone call was initiated at the NFI and received by a Vox IP Phone 4018 in Utrecht. Telephone recordings were made using a Marantz

[2] Refer to Appendix A for an orthographic transcription of the text.

professional PMD 661 portable recorder, which was connected to the telephone via a JK Audio Broadcast Host. Both studio and telephone speech samples were recorded at a sample frequency of 48 kHz and saved as 24 bits WAV files.

2.1.3. Procedure and data analysis

All read speech samples were saved as separate WAV files, resulting in 16 files (2 versions x 2 types of recording x 4 speakers), and automatically annotated at sentence, phrase, word and segment level in Praat (version 5.3.35; Boersma & Weenink, 2012). The automatic annotation was manually checked at word and segment level and corrected if necessary. For each sample, all vocalic information was extracted and saved as a separate WAV file. Vocalic information was also extracted by vowel and saved as separate WAV files for by-vowel LTF analysis. For all samples, formant settings in Praat were set to estimating 3 formants with a maximum of 3500 Hz . Formant measures were taken using Praat. Mean, standard deviation and 95% CI of LTFD[2,3] were computed for each sample, as well as LFTD[2] for sets of four back vowels (/ɔ, o:, u, ɑ/) and four front vowels (/i:, e:, ɪ, ɛ/). Kolmogorov-Smirnov Z was computed for pairs of LTFD[2,3] between and within speakers, and between types of recording (i.e. *studio* versus *telephone* recording). Kolmogorov-Smirnov Z is a measure of the distance between pairs of LTF distributions. Within-speaker comparisons compare LTFD[2,3] of the two versions of the text read by the same speaker; between-speaker comparisons generalize over both versions and measure the distance between both versions of the text (pooled) read by two speakers.

In telephone recorded samples, cases for which F3 was larger than 3000 Hz were excluded from analysis, in order to avoid interference of the expected effects of telephone transmission on frequencies above approximately 3,000 Hz (cf. Künzel, 2001). Also, cases for which the bandwidth of F2 and/or F3 exceeded 1000 Hz were excluded from analysis in both studio recorded and telephone recorded samples. In total, 23.5% (6,435 out of 27,395) of all cases were excluded from analysis.

2.2. Perception experiment

2.2.1. Participants

Four native speakers of Dutch took part in the perception experiment: DB, HB, HS, and JB.

Three participants were male (DB, HS, JB), one was female (HB). The participants were aged between 18 and 45. All listeners were assumed to have considerable experience listening to and distinguishing between the voices of MB and PB, since they are closely related to the speakers (i.e. father, mother, or brother of MB and PB, and fiancé of PB, respectively).

2.2.2. Materials

In order to restrict the number of stimuli, the data from one speaker (FW, whose geographical background deviates from that of the other speakers) were excluded from the perception experiment. The intensity of the other speech samples was normalized to 70 dB, in order to control for intensity as a cue in the speaker identification process. From each sample, 40 unique words were extracted and saved as separate WAV files, resulting in 480 stimuli (2 versions x 2 types of recording x 3 speakers x 40 words). Selection of words was based on word category (lexical versus functional category words), and number of syllables per word. Disyllabic, three-syllabic and four-syllabic lexical words were included first (cf. Bricker & Pruzansky, 1966), then supplemented with multisyllabic functional words and monosyllabic lexical and functional words. Table 1 displays word category and number of syllables per word of all stimuli which were included in the perception experiment.

Table 1
Overview of stimuli by number of syllable and word category.

	Number of syllables per stimulus					
	1		**2**		**>3**	
	N	**words**	**N**	**words**	**N**	**words**
functional	2	was, toen	2	voorbij, dichter	3	tenslotte, vervolgens, onmiddellijk
lexical	12	vraag, juist, kwam, jas, zijn, macht, hoe, blies, trok, gaf, slechts, zon	17	hadden, tweeën, sterkste, iemand, dikke, warme, aanhad, spraken, krijgen, trekken, begon, alle, blazen, harder, krachtig, stralen, daarop	4	noordenwind, discussie, voorbijganger, beamen

2.2.3. Procedure and data analysis

The listening procedure consisted of two separate experiments 1 and 2 and was preceded by a short practice session with samples from three speakers who were not included in experiments 1 and 2. Experiment 1 only contained stimuli extracted from studio recordings (N = 240); stimuli extracted from telephone recordings (N = 240) were included in experiment 2. Participants were assigned to the pair of experiments in a 2 x 2 between subjects factorial design: HS and JB participated first in experiment 1, and then in experiment 2; DB and HB participated first in experiment 2 and then in experiment 1. During the experiments, participants were presented with a random sequence of stimuli and instructed to decide on speaker identity by clicking a button on the screen. The names of MB, AK, and PB were presented on the buttons (in the aforementioned order, from left to right). Participants were told that there was no reaction time limit, and that there would not be a possibility to replay the stimuli. Stimuli were played from an ASUS K50IJ-SX laptop, which was connected to Sennheiser HD 477 headphones. There was a short pause after every 60 stimuli, and a longer pause between experiment 1 and 2. In total, participants took approximately 40 minutes to complete both experiments. All experiments were conducted in quiet rooms at the participants' homes.

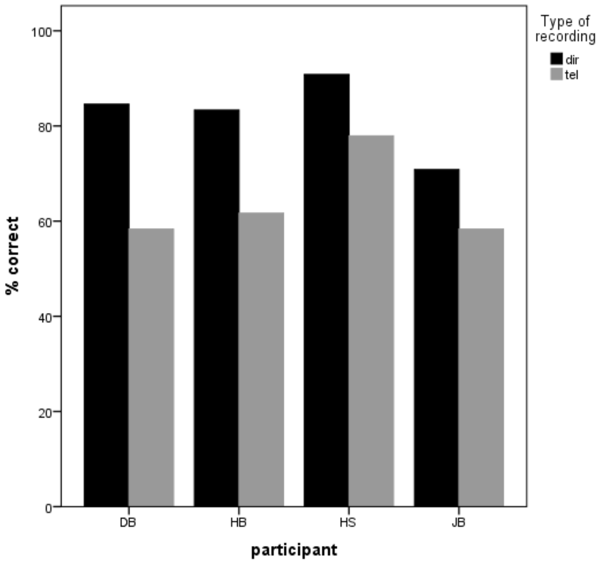
For every stimulus, expected response and given response were recorded. The amount of correct identifications per participant per speaker was analyzed by comparing scores between speech conditions (i.e. *studio recording* versus *telephone recording*) and by comparing listener confusion between the different speakers within each speech condition. Additionally, a Pearson's chi-square test of contingencies was used to evaluate whether success rate of auditory speaker identification was related to transmission characteristics of studio versus telephone recording. No cases were excluded from analysis.

3. Results and analysis

3.1. Influence of direct versus telephone recording on auditory speaker identification.

In order to evaluate whether success rate of auditory speaker identification depends on transmission characteristics of *telephone* versus *direct* recording, mean percentages of correct responses per speaker per type of recording are presented in Figure 1. For all participants in the perception experiment, a higher percentage of correct responses was found in directly recorded speech than in telephone transmitted speech.

Mean percentages of correct responses for twin versus non-twin speakers seem to show a similar pattern. As illustrated in Figure 2, directly recorded items were more likely to elicit a correct response from the listener than telephone recorded items for both twin and non-twin speakers. Importantly, this suggests that the influence of recording type on success rate of speaker identification was not different between twin versus non-twin speakers.



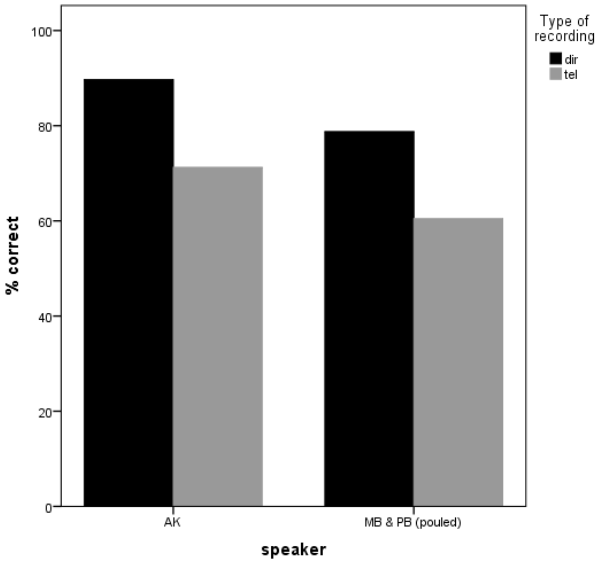


Figure 1. Mean percentage correct responses to directly recorded (*dir*) and telephone recorded (*tel*) items, per participant.

Figure 2. Mean percentage correct responses per speaker (AK, non-twin; versus MB and PB pooled, twins) per type of recording.

Confusion between speakers is further explored by providing frequencies and percentages of correct responses by speaker. Table 2 displays the confusion matrix for all speakers. From these data, it can be seen that AK seemed to be more often confused with PB than with MB; that MB seemed to be more often confused with PB than with AK; and that PB seemed to be more often confused with AK than with MB. Overall, 80.4% of all stimuli by AK were correctly identified; of all stimuli by MB and PB, 75.6% and 63.6% elicited correct responses, respectively.

Table 2

Overview of frequency count and percentages of responses by speaker per response category.

		response N (%)			
		AK	**MB**	**PB**	**total**
speaker	**AK**	515 (80.4%)	21 (3.3%)	104 (16.3%)	640 (100.0%)
	MB	51 (8.0%)	484 (75.6%)	105 (16.4%)	640 (100.0%)
	PB	163 (25.5%)	70 (10.9%)	407 (63.6%)	640 (100.0%)
		729	575	616	

3.2. Applicability and effectiveness of LTFD[2,3].

To give a general overview of mean LTFD[2,3] per speaker[3], mean LTFD2 and mean LTFD3 are presented in Figures 3 and 4, respectively. The error bar plots show that differences in LTF[2,3] means are not fully comparable for all speakers. Generally, for all speakers, SD of the mean is lower for mean F3 than for mean F2.

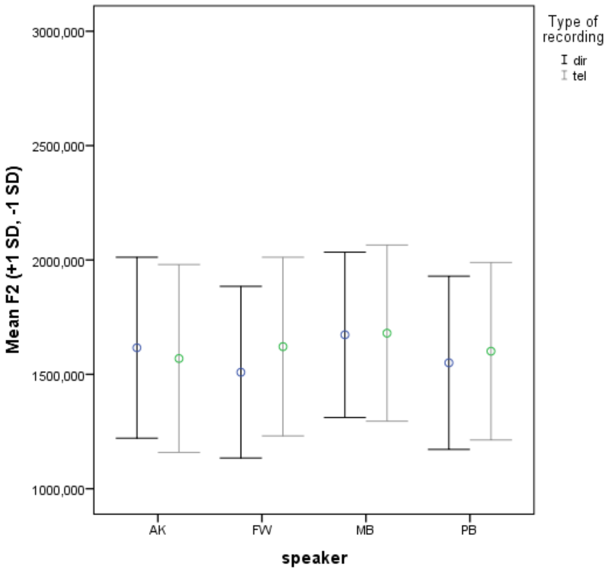


Figure 3. Mean LTF2 per type of recording per speaker. Horizontal lines represent the value of mean F2 ± 1 standard deviation.

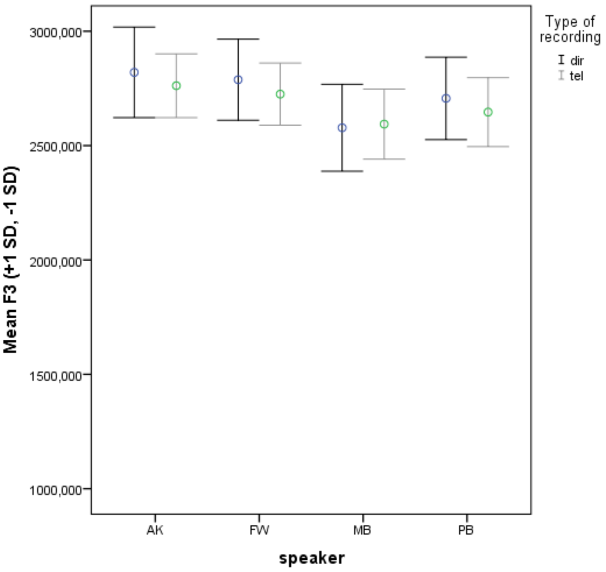


Figure 4. Mean LTF3 per type of recording per speaker. Horizontal lines represent the value of mean F3 ± 1 standard deviation.

Kolmogorov-Smirnov Z was computed for pairs of LTFD[2,3] between speakers. All Kolmogorov-Smirnov Z values for LTFD[2,3] for all pairs of speakers were significant at the .05 level, as is reported in Table C (Appendix C). An independent samples *t* test was used to compare the

[3] For a complete overview of mean, standard deviation and 95% CI of LTFD[2,3] of all speakers, refer to Appendix B.

average Kolmogorov-Smirnov Z values for distances between speakers in direct recordings to the average Kolmogorov-Smirnov Z values for distances in telephone recordings. The *t* test was not statistically significant, *t*(22) = -0.052, *p* = .959, and therefore it can be concluded that there was no statistical difference between the two categories of samples. Thus, the *t* test indicated that between-speaker distances in telephone recording and direct recording were similar.

An independent samples *t* test was used to compare the average Kolmogorov-Smirnov Z values for LTFD2 distances between speakers to the average Kolmogorov-Smirnov Z values for LTFD3 distances between speakers (see Appendix C). The *t* test was statistically significant, *t*(22) = -3.556, *p* = .002, two-tailed, *d* = 1.45; Kolmogorov-Smirnov Z differences for LTFD3 (*M* = 9.67, *SD* = 4.38) were on average 4.91 (95% CI [7.77, 2.05]) higher than differences for LTFD2 (*M* = 4.77, *SD* = 1.92). The value of Cohen's *d* (*d* = 1.45) indicates that 1.45 SD separates the Kolmogorov-Smirnov Z mean of LTFD2 distances from the Kolmogorov-Smirnov Z mean of LTFD3 distances; thus, the effect size of LTFD2 versus LTFD3 is large.

In order to investigate whether LTFD[2,3] distances within speakers are smaller than distances between twins within a monozygotic twin pair, and whether LTFD[2,3] between speakers of non-twin pairs are larger than within-speaker and within-twin pair distances, Kolmogorov-Smirnov values were also computed for between-version LTFD[2,3] differences within speakers (see Appendix C for all values). All distances between both twin and non-twin pairs are significant at the .05 level. Four within-speaker distances are also significant at 0.05 level, twelve are not. FW is the only speaker for whom none of the within-speaker distances are significant. A Mann-Whitney *U* test suggested that the Kolmogorov-Smirnov Z values of the non-twin pairs (*Mean Rank* = 12.40, *n* = 20) were not significantly different from the Kolmogorov-Smirnov Z values of the twin pair (*Mean Rank* = 13.00, *n* = 4), *U* = 38.00, *z* = -0.16 (not corrected for ties), *p* = .877, two-tailed. Although the Mann-Whitney *U* should be interpreted cautiously because the sample size is very small, the absence of a categorical difference between LTFD[2,3] distances between twin versus non-twin speakers seems to be reflected also in Table 3, which includes Kolmogorov-Smirnov Z values and SD of distances between twin versus non-twin pairs of speakers.

Table 3

Kolmogorov-Smirnov Z values and standard deviations of LTFD[2,3] distances between twin versus non-twin speaker pairs per type of recording per formant.

		non-twin pairs			twin pairs		
		N	**Mean K-S Z**	**SD**	**N**	**K-S Z**	**SD**
dir	**F2**	5	5.03	2.34	1	4.52	n.a.
	F3	5	9.08	5.33	1	11.00	n.a.
tel	**F2**	5	4.32	1.99	1	5.89	n.a.
	F3	5	10.67	4.41	1	6.30	n.a.

3.3. LTFD2 in sets of front and back vowels.

In order to explore to what extent it is possible to distinguish between speakers by investigating vowel subsets of front versus back vowels, all front and back vowels of all speakers in directly recorded speech were plotted in Figure 5.[4] The figures show that visual differences between speakers are relatively large in the vowels /e:, i:, u/, and relatively small in the vowels /ɪ, ɛ, ɑ/. Apparently, Figure 5 shows no categorical difference in speaker specificity between back versus front vowels. Rather, from these data, specific individual vowels from both categories seem to be relatively speaker-specific compared to other vowels from both categories.

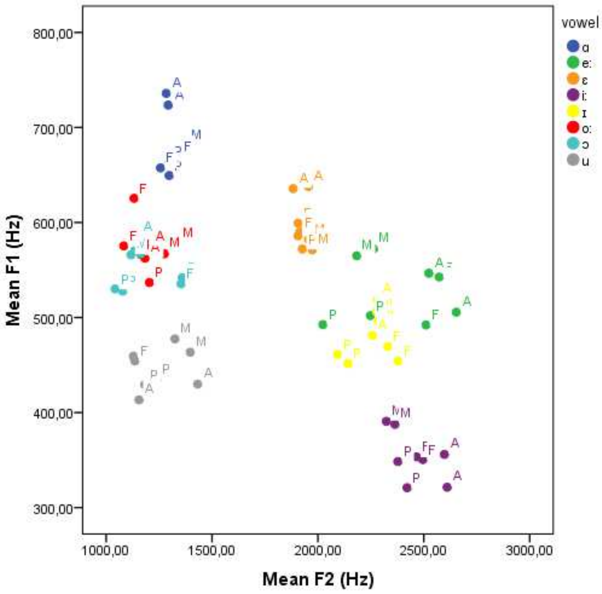


Figure 5. Vowel space of back (/ɔ, o:, u, ɑ/) and front (/i:, e:, ɪ, ɛ/) vowels for speakers AK (A), FW (F), MB (M) and PB (P) in direct recording.

[4] Other than usual, the x-axis and y-axis of Figure 7 are scaled ascendingly. The present vowel space is therefore a mirror image of a conventional vowel space.

Further analysis of the distribution of individual vowels in the vowel space was conducted by plotting all vowels of all speakers in Figure D (see Appendix D). Since the position of vowels in the vowel space is seriously affected by telephone transmission of the speech signal, Figure D contains data from directly recorded samples only. Although the relative positions of some individual vowels in the vowel spaces of MB and PB are similar to those in the vowel spaces of AK and FW, the most striking similarity between MB and PB (as opposed to AK and FW) is the relatively dense area between [1750-2400, 450-600] (F2,F1 in Hz), in which the vowels /e:, ɛ, ɪ, ɵ, ə/) are positioned. Also, Figure 5 reveals that the F2 range of MB and PB is smaller than the F2 range of AK and FW. This is particularly reflected in the F2 value of the most extreme vowel on the x-axis, which is below 2,500 Hz for MB and PB, but more than 2,500 Hz for AK and FW. Figure 5 thus seems to suggest that maximum, minimum and range values of LTF2 per speaker might contribute to speaker-specificity in vowel space size and shape.

4. Discussion and conclusion

By means of a small-scale perception experiment and acoustic analysis of speech samples of four speakers, the present study evaluates the effectiveness of LTF(D)[2,3] as a measure in speaker comparison. The following section provides a discussion of the results and conclusions per research question.

First of all, confusion between speakers was expected to be higher in telephone speech relative to studio speech. In the perception experiment, stimuli from directly recorded samples were more likely to elicit correct responses from listeners than telephone recorded items. The abovementioned effect of recording type was not different between twin versus non-twin speakers. Transmission of directly recorded speech via a telephone connection thus indeed influenced the speech signal in such a way that perceptional speaker identification was impeded.

Secondly, confusion between twin speakers was hypothesized to be higher than confusion between non-twin speakers. Overall, stimuli by AK seemed to be more often identified correctly than stimuli by MB and PB (80.4%, 75.6% and 63.6%, respectively). Although the listeners did not have as much experience with AK's voice as with the voices of MB and PB, the fact that she was a non-twin speaker while MB and PB were twin speakers presumably caused the listeners to be able to identify stimuli by AK more successfully. However, contrary to the hypothesis, PB was more often confused with non-related speaker AK (25.5%) than with her twin sister MB (10.9%), while MB was more often confused with PB (16.4%) than with AK (8.0%). There is no straightforward explanation for this apparent asymmetry between confusion data from MB and PB. However, most importantly, these data suggest that confusion between twin's voices and non-related speakers should not always be assumed to be symmetrical.

In line with the hypotheses, no difference was found between distances between speakers in direct recording compared to telephone recording. Also, as expected, no categorical difference was found between mean F2 or F3 in directly versus telephone recorded samples. However, Kolmogorov-Smirnov Z distances between speakers *were* on average higher for LTFD3 than for LTFD2. This is congruent with findings by Moos (2010), and Gold et al. (2013), who suggest that LTFD3 is a more effective speaker discriminant measure than LTFD2.

All LTFD[2,3] distances between both twin and non-twin pairs of speakers were significant at the .05 level. Although most of the within-speaker distances between LTFD[2,3] were not significant, some within-speaker distances between distributions *were* significant, possibly due to habituation effects. Most of the within-speaker differences between distributions were found in pairs of LTF3 distributions from directly recorded samples. Since LTFD3 was previously found to be a more sensitive measure of speaker-specificity, and since direct recordings are assumed to contain more speaker-specific perceptual information than telephone recordings, the occurrence of significant LTFD within-speaker distances in those pairs is not completely arbitrary: based on the previously described results, if a within-speaker difference occurs, it should be expected in measurements of the most sensitive parameter (i.e. LTFD3), and in the most informative condition (i.e. direct recording). Thus, even though LTFD[2,3] measures seem to be successful in distinguishing between speakers, caution should be taken to generalize over performance of (specifically) LTFD[3] in direct recordings. Further research is needed to explore the power of LTFD[2,3] in capturing within-speaker differences.

Contrary to the hypothesis, the present data suggest that there was no categorical difference in LTFD[2,3] distances between twin versus non-twin pairs of speakers. This finding was further supported by data from the perception experiment, which suggested that speech of a twin speaker was not always perceived as being more closely related to speech of the other twin speaker than to speech of a non-related speaker.[5] Following Loakes (2006), the absence of a categorical difference between twin versus non-twin pairs of speakers might be due to a relatively high degree of differences in the present monozygotic twin pair for this parameter (i.e. LTFD[2,3]). In that case, further research with other monozygotic twin pairs and comparable non-related speakers should be conducted to shed more light on the use of LTFD[2,3] as a measure in speaker comparison. Alternatively, further research could also investigate the degree of differences in the present monozygotic twin pair for parameters other than LTFD[2,3].

A separate analysis of by-vowel LTFD2 in front vowels was expected to be a more effective speaker discriminant measure than LTFD[2,3] analysis across vowels (cf. Loakes, 2004), given the generalizing quality of LTFs across vowels (as opposed to by-vowel LTFs). By-vowel analysis in direct recording showed no categorical differences in speaker specificity between back versus front vowels, but indicated that specific individual vowels from both categories seemed to be relatively speaker-specific compared to other vowels. Furthermore, although plots of by-vowel LTF measurements did not reveal straightforward differences between speakers, the data suggested they might still be helpful for distinguishing between speakers. Speaker-specific information regarding the size of the vowel space (e.g. range of mean F[1,2], and/or the position of specific vowels, e.g., /i, u, a:/), and the distribution of vowels within the vowel space can indeed be illustrative of speaker-specific differences leveled out previously by across-vowel LTFD[2,3] measurements. In this regard, differences between speakers with different dialectal and/or language backgrounds should be interpreted cautiously, since the location of individual vowels in the vowel space is partly determined also by a language-specific base-of-articulation property (Bradlow, 1995). As Bradlow (1995) suggests that tightness of within-category clustering of vowels in the vowel space might not be language-specific, the focus of further research should probably indeed be on by-vowel LTF analysis rather than on the relative distribution of *all* vowels in the vowel space.

[5] It should be noted, however, that the absence of a *categorical* acoustic difference in LTFD[2,3] between twin- versus non-twin speakers is *not* reflected in the perception data, where AK (as mentioned previously) was successfully identified significantly more often than MB and PB (pouled, see Figure 2).

Generally speaking, telephone transmission of speech samples negatively influenced speaker identification in both twin and non-twin pairs of speakers. Also, the data suggested that confusion between twin’s voices and non-related speakers should not always be assumed to be symmetrical. Overall, LTFD3 seems to be a more effective speaker discriminant parameter than LTFD2 in both telephone recorded and directly recorded samples. Both LTFD2 and LTFD3 measurements effectively discriminated between all pairs of speakers, although the occurrence of within-speakers differences between distributions emphasizes the need for further research into the use of LTFD[2,3] as a measure in speaker comparison. Furthermore, the results suggest that analyses of individual vowel spaces in directly recorded samples (but presumably *not* in telephone recordings, since F1 is severely affected by telephone transmission; cf. Künzel, 2001) could be a useful addition to the arsenal of speaker discriminant measures in speaker comparison studies.

Acknowledgements

I very gratefully acknowledge the many ideas, comments, and inspiring support of my first supervisor, dr. Willemijn Heeren. I also thank AK, FW, MB, DB, HB and JB for participating in the study; the NFI for their ideas and suggestions as well as their practical support in recording the telephone-transmitted speech; and dr. Hugo Quené for his helpful comments on the research proposal. I especially thank HS for his participation in the experiment and for his invaluable support throughout the writing process.

5. References

Al-Tamimi, J. E. & Ferragne, E. (2005). Does vowel space depend on language vowel inventories? Evidence from two Arabic dialects and French. *INTERSPEECH*, 2465-2468.

Becker, T., Jessen, M., & Grigoras, C. (2008). Forensic speaker verification using formant features and Gaussian Mixture Models. *INTERSPEECH,* 1505-1508.

Bradlow, A. (1995). A comparative acoustic study of English and Spanish vowels. *Journal of the Acoustical Society of America, 97*(3), 1916-1924.

Bricker, P.D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *Journal of the Acoustical Society of America, 40*(6), 1441-1449.

Broersma, P., & Weenink, D. (2012). Praat 5.3.35. [retrieved on 07/12/2013 from www.praat.org]

Byrne, C. & Foulkes, P. (2004). The 'mobile phone effect' on vowel formants. *Speech, Language and the Law, 11*(1), 83-102.

Engstrand, O., & Krull, D. (1991). Effect of inventory size on the distribution of vowels in the formant space: preliminary data for seven languages. *PERILUS*, *13*, 15-18.

Gold, E. & French, P. (2011). International practices in forensic speaker comparison. *International Journals of Speech, Language and the Law*, *18*(2), 293-307.

Gold, E., French, P. & Harrison, P. (2013). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. *Proceedings of Meetings on Acoustics, 19,* 1-7.

Gonzalez, J., & Carpi, A. (2004). Early effects of smoking on the voice: A multidimensional study. *Medical Science Monitor*, *10*(12), CR649-CR656.

Gussenhoven, C. (1992). Dutch. *Journal of the International Phonetic Association, 22*(1-2), 45-47.

Jessen, M., & Becker, T. (2010). Long-term formant distribution as a forensic-phonetic feature. *The Journal of the Acoustical Society of America, 128*(4), 2378-2378.

Künzel, H. J. (2001). 'Beware of the telephone effect': the influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics, 8*(1), 80-99.

Loakes, D. (2004). Front vowels as speaker-specific: Some evidence from Australian English. In *Proceedings of the Australian International Conference on Speech Science* (289-294).

Loakes, D. (2006). Variation in long-term fundamental frequency: Measurements from vocalic segments in twins' speech. In *Proceedings of the 11th Australian International Conference on Speech Science & Technology* (205-210).

Maddieson, I. (1984). *Patterns of Sounds*. Cambridge: Cambridge University Press.

Moos, A. (2010). Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech. *The Phonetician, 101*, 7-24.

Nolan, F. & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *Journal of Speech, Language and the Law, 12*, 143-173.

Nolan, F. & Oh, T. (1998). Identical twins, different voices. *Forensic Linguistics, 3,* 39-49.

Rose, P.J. (2003). The technical comparison of forensic voice samples. In Freckelton, I., & Selby, H. (Eds.), *Expert Evidence* (chapter 99). Sydney: Thomson Lawbook Company.

San Segundo Fernández, E. (2013). A phonetic corpus of Spanish male twins and siblings: Corpus design and forensic application. *Procedia Social and Behavioral Sciences*, *95*, 59-67.

San Segundo Fernández, E. (2014). Forensic speaker comparison of Spanish twins and non-twin siblings: A phonetic-acoustic analysis of formant trajectories in vocalic sequences, glottal source parameters and cepstral characteristics. Doctoral Thesis, UIMP.

Stevens, K. (1989). On the quantal nature of speech. *Journal of Phonetics, 17,* 3-46.

6. Appendices

Appendix A: orthographic transcription of recorded passage (Gussenhoven, 1992)

De noordenwind en de zon hadden een discussie over de vraag wie van hun tweeën de sterkste was, toen er juist iemand voorbij kwam die een dikke, warme jas aanhad. Ze spraken af dat wie de voorbijganger ertoe zou krijgen zijn jas uit te trekken de sterkste zou zijn. De noordenwind begon uit alle macht te blazen, maar hoe harder hij blies, des te dichter de voorbijganger zijn jas om zich heen trok. Tenslotte gaf de noordenwind het maar op. Vervolgens begon de zon krachtig te stralen, en onmiddellijk daarop trok de voorbijganger zijn jas uit. De noordenwind kon toen slechts beamen dat de zon de sterkste was.

Appendix B: Mean, SD and 95% CI of F2 per sample per speaker.

Table B1
Mean, SD and 95% CI per sample for speaker MB.

		MB			
		N	**Mean (Hz)**	**SD (Hz)**	**95% CI (Hz)**
dir	F2	2876	1672.91	361.46	[1659.69, 1686.12]
	F3	2876	2578.12	190.05	[2571.17, 2585.07]
tel	F2	3403	1680.16	385.12	[1667.21, 1693.10]
	F3	3403	2594.01	153.03	[2588.86, 2599.15]

Table B2
Mean, SD and 95% CI per sample for speaker AK.

		AK			
		N	**Mean (Hz)**	**SD (Hz)**	**95% CI (Hz)**
dir	F2	2129	1616.34	395.30	[1599.54, 1633.14]
	F3	2129	2820.14	197.95	[2811.73, 2828.56]
tel	F2	2151	1569.21	410.77	[1551.84, 1586.58]
	F3	2151	2761.80	139.43	[2755.90, 2767.69]

Table B3
Mean, SD and 95% CI per sample for speaker PB.

		PB			
		N	**Mean (Hz)**	**SD (Hz)**	**95% CI (Hz)**
dir	F2	2104	1550.50	379.00	[1534.29, 1566.70]
	F3	2104	2706.58	180.15	[2698.87, 2714.28]
tel	F2	2797	1601.20	387.78	[1586.82, 1615.58]
	F3	2797	2646.44	150.99	[2640.85, 2652.04]

Table B4
Mean, SD and 95% CI per sample for speaker FW.

		FW			
		N	**Mean (Hz)**	**SD (Hz)**	**95% CI (Hz)**
dir	F2	2670	1509.24	375.15	[1495.00, 1523.47]
	F3	2670	2788.29	177.12	[2781.57, 2795.01]
tel	F2	2830	1621.24	390.54	[1606.85, 1635.64]
	F3	2830	2725.06	135.76	[2720.06, 2730.07]

Appendix C: Kolmogorov-Smirnov Z values for LTFD[2,3] per speaker pair.

Table C

Kolmogorov-Smirnov Z for between-speaker comparison of LTFD2 and LTFD3. * indicates significant values with p < 0.05. Distances within speakers (between versions) are in grey; distances between speakers of non-twin pairs are in white; distances between speakers of twin pairs are in bold.

			Speaker			
			AK	**FW**	**MB**	**PB**
AK	**F2**	**dir**	K-S Z = 1.060, p = 0.211	K-S Z = 3.996, p < .001*	K-S Z = 5.176, p < .001*	K-S Z = 2.797, p < .001*
		tel	K-S Z = 1.043, p = 0.227	K-S Z = 2.359, p < .001*	K-S Z = 6.782, p < .001*	K-S Z = 3.985, p < .001*
	F3	**dir**	K-S Z = 1.975, p = 0.001*	K-S Z = 2.581, p < .001*	K-S Z = 15.012, p < .001*	K-S Z = 7.340, p < .001*
		tel	K-S Z = 1.202, p = 0.111	K-S Z = 5.097, p < .001*	K-S Z = 15.621, p < .001*	K-S Z = 10.690, p < .001*
FW	**F2**	**dir**		K-S Z = 1.059, p = 0.212	K-S Z = 8.939, p < .001*	K-S Z = 4.242, p < .001*
		tel		K-S Z = 1.343, p = 0.054	K-S Z = 5.943, p < .001*	K-S Z = 2.547, p < .001*
	F3	**dir**		K-S Z = 1.033, p = 0.236	K-S Z = 14.137, p < .001*	K-S Z = 6.345, p < .001*
		tel		K-S Z = 1.103, p = 0.175	K-S Z = 14.282, p < .001*	K-S Z = 7.670, p < .001*
MB	**F2**	**dir**			K-S Z = 0.796, p = 0.550	**K-S Z = 4.524, p < .001***
		tel			K-S Z = 0.814, p = 0.521	**K-S Z = 5.891, p < .001***
	F3	**dir**			K-S Z = 2.676, p < .001*	**K-S Z = 10.998, p < .001***
		tel			K-S Z = 2.040, p < .001*	**K-S Z = 6.303, p < .001***
PB	**F2**	**dir**				K-S Z = 1.999, p = 0.001*
		tel				K-S Z = 1.250, p = 0.088
	F3	**dir**				K-S Z = 0.835, p = 0.489
		tel				K-S Z = 1.121, p = 0.162

Appendix D: Vowel spaces in direct recording per speaker.

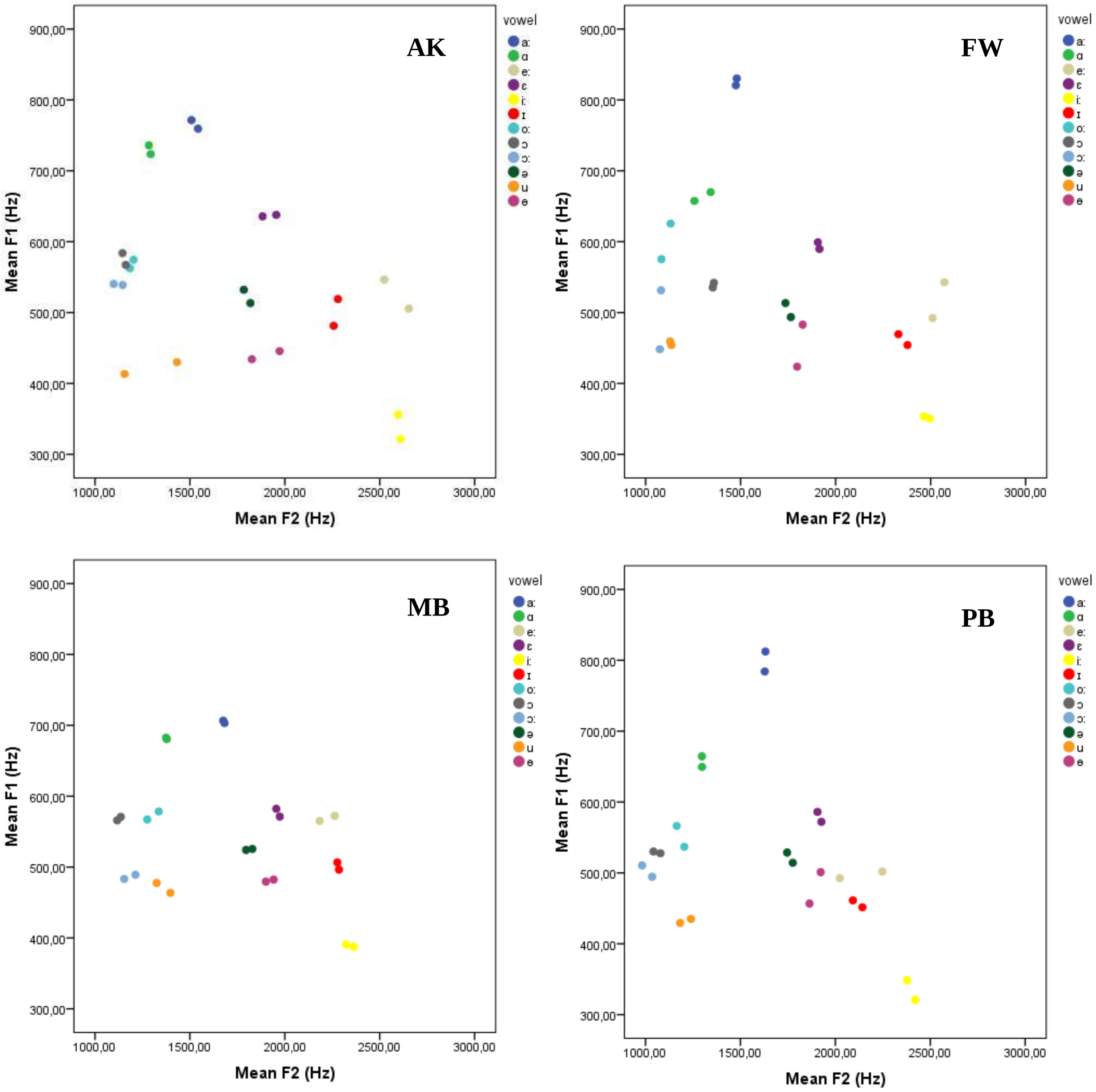


Figure D. Vowel space of speakers AK, FW, MB and PB in direct recording.