POSTPRINT VERSION of article published in IJSLL 27.1 2020, DOI: 10.1558/ijsll.41058

The Contribution of Dynamic versus Static Formant Information in Conversational Speech

Abstract
The relative contributions of static and dynamic formant representations to speaker-specificity were investigated in conversational speech and in two vowels varying in inherent spectral change. Using polynomial fits, the contribution of dynamic formant coefficients to speaker-specificity relative to that of the formant intercept was investigated in the diphthongal vowel [ei] taken from English and Dutch conversational speech. The [ei] tokens were sampled from various linguistic contexts and analysed in an LR approach. Results show that formant dynamics contain speaker-specific information in conversational speech even though the high contextual variation seems to reduce its effect relative to that reported by earlier work.

Vowels differ in inherent dynamicity, and therefore the added value of dynamic formant information to speaker-specificity was also compared between vowels differing in inherent spectral change. Using Dutch data, the contribution of formant dynamics to speaker-specificity was compared between [ei] and [a:] tokens produced by the same speakers. Formant dynamics in conversational speech only contributed to speaker-specificity in the diphthong [ei], not in the monophthong [a:].

Keywords
speaker-specificity, formant dynamics, conversational speech, vowel-inherent spectral change

Bio
Willemijn Heeren is an associate professor in phonetics at the Leiden University Centre for Linguistics and the principal investigator of a research project on the interaction between indexical and linguistic information in speech (2017-2022, NWO VIDI nr. 276-75-010). She is also a forensic speech scientist at the Netherlands Forensic Institute (NFI). Her research interests include forensic phonetics, socio-phonetics, experimental linguistics and language learning.

The Contribution of Dynamic versus Static Formant Information in Conversational Speech

1. Introduction

Traditionally, vowels have been phonetically characterised using formant values measured at their temporal midpoint, but it is well-known that vowel formants often change in the course of a vowel's pronunciation. This spectral change is not only the result of co-articulation with neighboring sounds. It is also inherent to the pronunciation of vowels when produced in isolation, in both diphthongs and monophthongs (e.g. Nearey and Assman 1986). This latter type of inherent variation is referred to as vowel-inherent spectral change. It contributes to vowel recognition (e.g. Nearey and Assman 1986), especially when the combination of onset plus offset information are presented rather than the mid-vowel section (Morrison and Nearey 2007). Just as mean formant values show by-speaker variation (cf. Peterson and Barney 1952), vowel-inherent spectral change does. As will be discussed further in section 1.1, many studies have demonstrated the value of dynamic formant representations for speaker classification and discrimination (e.g. McDougall 2006). This paper builds on that work by extending its evidence base in conversational speech styles, but also by directly comparing dynamic with static formant representations to thus study the relative contribution made by the dynamic information per se. It will do so in a diphthongal and a monophthongal vowel.

Even though earlier research shows an important contribution of dynamic formant representations to speaker classification and discrimination (e.g. Ingram et al. 1996; McDougall 2004, 2006; McDougall and Nolan 2007; Morrison 2008, 2009a; Thaitechawat and Foulkes 2011; Fejlovà et al. 2013; Zuo and Mok 2015), this body of work has some limitations. First, most vocalic materials studied so far were diphthongs, as a result of which the usefulness of formant dynamics for speaker classification from monophthongs remains understudied. The literature on vowel-inherent spectral change, however, would suggest that also monophthongs may be expected to contain speaker-dependent differences in their dynamics. There are a few studies that included such vowels (Fejlovà et al. 2013; Rose 2015; Hughes et al. 2016; Wang et al. 2019), but results varied (details in section 1.1).

Second, in the large majority of the work on formant dynamics no distinction was made between the contributions of static versus dynamic formant information to speaker-specificity (but see Hughes et al. 2016): most formant representations used in studies on dynamics, however, included static or average formant information. Not all speaker-discriminatory potential can therefore be attributed to dynamic information. The general lack of differentiation between these information sources holds for both acoustic-phonetic investigations and studies that took automatic speaker recognition (ASR) approaches to evaluate the speaker-specificity of formants (e.g. Zhang et al. 2013; Franco-Pedroso and Gonzalez-Rodriguez 2016).

Finally, much of the speech materials used in earlier phonetic-acoustic, i.e. non-ASR, approaches to formant dynamics consisted of read speech or controlled lab speech (but see Hughes et al. 2016). This is only partially representative of the spontaneous conversation typical of everyday communication and of the police interviews or tapped telephone conversations often encountered in forensic speaker comparisons. These relatively high levels of control in the speech data used so far, in combination with the inclusion of static information in formant representations, may have resulted in an over-estimation of formant dynamics' relevance.

The current investigation aimed to address these issues by (1) assessing the relative contributions of static and dynamic formant information to speaker-specificity in an inherently-dynamic, diphthongal vowel in conversational speech styles, and by (2) comparing the relative contributions of formant dynamics to speaker-specificity between a diphthongal and a monophthongal vowel from Dutch conversational speech.

1.1 Formants and their dynamics as a speaker-dependent feature

The classical problem of invariance in speech finds its background in the variability that speech signals contain, originating from contextual as well as within- and between-speaker variation (see e.g. Blumstein and Stevens 1981, for a discussion). For retrieval of the linguistic message from speech, variation has been considered problematic. However, for retrieval of speaker-dependent information in speech, variation is essential. Speakers are assumed to have unique combinations of vocal tract anatomy and use of those structures during articulation (e.g. Johnson, Ladefoged and Lindau 1993; Weirich and Simpson 2018). Moreover, an individual's language knowledge adds to this combination, co-determining the use of the biologically given vocal tract. When looking at vowels specifically, individual speakers' formant frequencies are reflections of the speaker's vocal tract and articulation, combined with their knowledge of a particular language (variety)'s sound system. In spite of the assumed individuality, there is much within-speaker variation in acoustic vowel realizations caused by speech style, emotion, addressee etc. This within-speaker variation also results in overlap between different speakers' realizations of the same (or even different) vowels. Moreover, further between-speaker overlap is found through shared varieties, such as dialects and ethnolects.

Earlier work on the speaker-dependency of vowels showed that they contain relevant information in the durational (Van de Heuvel 1996) and spectral domains (Van de Heuvel 1996; Rose 2002: 230-238), and illustrated that formant frequencies may reflect information about different speakers in casework (Nolan and Grigoras 2005). Long-term formant distributions have also been investigated and confirmed as containing speaker-specific information (Moos 2010; Gold 2014, ch. 5). Rose (1999) found that long-term within-speaker differences in formants produced by six male speakers were smaller than between-speaker differences. This combination of smaller within- and larger between-speaker variability supports the speaker-dependency hypothesis. A next step in speaker-dependent information in vowels was made by considering formant dynamics (see McDougall 2004, for an overview of early work). The reasoning behind this was that in addition to a reflection of a speaker's vocal tract resonances, i.e. mean formants, vowel pronunciation contains information on that speaker's articulation, i.e. formant dynamics.

The majority of acoustic-phonetic studies undertaken so far used read speech to assess the usefulness of dynamic formant representations for speaker classification (e.g. Ingram et al. 1996; McDougall 2004, 2006; McDougall and Nolan 2007; Morrison 2008, 2009a; Thaitechawat and Foulkes 2011; Fejlovà et al. 2013; Zuo and Mok 2015). These representations generally included both static and dynamic formant information. Moreover, in the course of this work several improvements in the modelling of formant tracks were made. An early study including 15 male speakers determined dissimilarities between time series of formant measurements, using samples from different recordings of the same speaker and recordings of different speakers (Ingram et al. 1996). Results showed that speaker classification based on these measurements was over 85% correct. This way of computing similarities between real-time formant tracks was followed up by the use of time-normalized series of measurements, which eliminated the problem of unequal vowel durations in comparisons. This approach was taken by Rose (1999), McDougall (2004), Thaitechawat and Foulkes (2011), and Zuo and Mok (2015). Using five male Australian English speakers producing /aɪ/ in C/aɪk/ contexts (e.g. *hike*, *like*), McDougall (2004) showed that time-normalized first through third formant (F1-F3) tracks yielded high classification performance. Five male speakers of Thai were classified correctly up to 100% using F1-F3 series from read diphthongs (Thaitechawat and Foulkes 2011).

The test case of identical twins, who were furthermore bilingual speakers of Shanghainese and Mandarin, was used by Zuo and Mok (2015) to investigate the role of the individual, not the anatomy, in read speech formant dynamics of /ua/. Using time-normalized formant tracks for F1-F3, they showed that twins were discriminated less well in their dominant language, but

that classification performance overall was still well-above chance level. Furthermore, in this study a comparison was made between time-normalized series of measurements and polynomial approximations of the formant contours; the fits' coefficients resulted in classification performance comparable to results with time series. Fitting the time-series with polynomial curves, however, solved the problem of correlation between sequential measurements in time-normalized series (McDougall 2006; McDougall and Nolan 2007; Morrison 2008, 2009a). McDougall (2006) re-analyzed the data from McDougall (2004) using curve fitting and showed that good performance was obtained with fewer parameters. Morrison (2009a) studied five diphthongs, added discrete cosine transforms as a fit method to capture dynamics, and demonstrated that the optimal LR system parameters varied with vowel. More recent studies, including those using ASR approaches, find that dynamic formant representations are a useful feature also in more spontaneous, conversational speech (e.g. Zhang et al. 2013; Franco-Pedroso and Gonzalez-Rodriguez 2016; Hughes et al. 2016; Enzinger and Morrison 2017). However, when using formant representations that include both static and dynamic information, the relative increase in speaker-dependent information through the addition of dynamic information remains unclear (but see Hughes et al. 2016).

In the majority of the aforementioned studies, diphthongs or segmental combinations were examined, i.e. intervals with relatively large inherent formant transitions (e.g. McDougall 2004; 2006; Rose 1999; McDougall and Nolan 2007; Morrison 2009a). In spontaneous speech, the variable phonetic contexts of the target speech sounds are likely to reduce the informational value of formant dynamics: dynamics are partly determined by neighboring sounds that differ between tokens and may be compromised by shortened durations. Moreover, monophthongal vowels –in spite of their inherent spectral change– may contain even less dynamic information in the case of non-controlled recording circumstances due to increased variation in utterance lengths. Longer utterances have increased speech rates and therefore shorter vowel durations, which results in more coarticulation as well as vowel reduction. These effects especially impact inherently shorter vowels, including monophthongs. There are a few studies addressing these concerns for forensic speaker comparisons. In Rose (2015), some contextual variation was included in the design when comparing mid-vowel measurements to formant trajectories of the steady-state vowel /ɜ/. In map task recordings, eight different word contexts were used to record the vowel. Results showed stronger speaker evidence with formant trajectories than mid-vowel measurements. Similar results were reported for Czech monophthongs taken from different consonantal contexts, and from stressed and unstressed syllables (Fejlovà et al. 2013). This study did not differentiate between conditions with and conditions without dynamic formant information, but when the order of the polynomial fit was increased from linear to cubic, speaker classification performance improved. In a study on the speaker-dependency of hesitation markers *uh* and *um* using spontaneous British English speech (i.e. with varying phonetic contexts for the hesitation markers), formant dynamics only improved classification for *um*, including the transition from vowel to nasal, but not for *uh*, without transition (Hughes et al. 2016).

In sum, formant dynamics seem to contain speaker-dependent information, but given the current state-of-the art, their added information value, especially in conversational speech, and the role of vowel characteristics therein, are understudied. The current investigation contributes to filling that gap.

1.2 The diphthongal and monophthongal vowels under study

As the current investigation is a follow-up to earlier work on the speaker-specificity of vowels in Dutch, the Dutch vowel inventory was taken as a starting point. Dutch has a relatively rich vowel inventory consisting of 13 monophthongs, lax /ɪ, ɛ, ɑ, ɔ, ʏ, ə/ and tense /i, y, eː, aː, øː, oː, u/, and three diphthongs /ɛi, œʏ, ɑu/ (Gussenhoven 1999). Some of these vowels can be

pronounced lengthened or nasalized when used in loan words, and /œː/ can also be found in loan words. To my knowledge, only the speaker-specificity of the Dutch corner vowels, /i, aː, u/, has been studied (Van den Heuvel 1996). This work showed that out of these three monophthongs, the inherently long /aː/ contained the most speaker-dependent information. A more recent study on spontaneous speech from a closely-related language, German, confirmed that /aː/ may be considered a relatively speaker-specific vowel (Schindler and Draxler 2013). Neither study, however, included dynamic formant representations. Against the background of these findings, the current study investigated the contribution of dynamic relative to static formant information to speaker-dependent information in /aː/, a phonologically monophthongal vowel. Moreover, this comparative analysis was also conducted for vowels containing more vowel-inherent change: diphthongs.

The Dutch vowel /eː/ is not counted as one of the language's phonological diphthongs (see above), but often shows much intrinsic spectral change from /e/ towards /i/ (e.g. Adank et al. 2007), making it a phonetic diphthong. This intrinsic change developed over the past century (Van de Velde 1996), but is contextually restricted: diphthongal realization is blocked before liquids and semi-vowels (cf. Voeten, in progress). In the present investigation, this phonetic diphthong was preferred over a phonological diphthong, because it allows for the comparison with the British English phonological diphthong /eɪ/ that is realized in a comparable way (in words such as *play*, *great*). The inclusion of phonetically similar vowels in two languages allows for repetition (i.e. replication) of one of the experiments in this study; the use of two independent datasets taken from two different languages is thought to support the generalization of this study's results. The choice for this diphthong furthermore ties in with existing work on the Australian English vowel /eɪ/ (Morrison, 2009a). In many varieties of English, the transition towards an /i/-like vowel is part of the vowel's phonemic characterization: /ei/ or /eɪ/ (e.g. Jones 1957: 100; Roach 2004).

To express the comparability in the phonetic realization of Dutch /eː/ and English /eɪ/ I will also refer to these vowels as [ei] for either language in the rest of this paper. In both languages, the F1 values of [ei] lower in the course of the vowel's pronunciation while for F2, there is an increase with time. F3 is expected to stay relatively stable through time.

1.3 Research questions

The first question is whether dynamic formant information in a diphthongal vowel contributes speaker-dependent information in conversational speech, in addition to static formant information. The question is addressed in Experiments 1a and 1b, once using tokens of the vowel [ei] sampled from a Dutch speech collection (1a) and once from an English one (1b). These collections both contain conversational speech, but also show differences, e.g. in the recording conditions and the level of spontaneity of the speech (see section 2.1). As mentioned in 1.2 this variation in languages and speech collections was intended to support the generalization of this study's results on the contribution of formant dynamics. The question was operationalized by using polynomial fits as formant representations, and differentiating between the static coefficient, i.e. the intercept, and the higher coefficients capturing dynamic information. The prediction was that in conversational speech, formant dynamics contribute speaker-dependent information to static formant information in an inherently-dynamic vowel.

The second question is whether the speaker-specificity of dynamic formant information in conversational speech varies with the degree of vowel-inherent spectral change; monophthongs show less inherent change than diphthongs do. This question was addressed in Experiment 2, using Dutch data only; the relative contribution of formant dynamics from [ei] realizations was compared to that from [aː] realizations produced by the same set of speakers. It was predicted that the contribution of formant dynamics is higher in the diphthongal vowel [ei] than in the monophthongal vowel [aː].

2. Experiment 1: The contribution of formant dynamics to speaker-specificity in conversational speech

2.1 Materials and speakers

For Dutch, spontaneous telephone conversations recorded in speakers' home environments from the Spoken Dutch Corpus were used (stereo recording, 8 kHz sampling frequency, see Oostdijk 2000). The full corpus consists of fifteen components, covering different speech styles such as read and conversational speech. The telephone speech was recorded from a home environment through a switchboard with a frequency pass band of 340-3,400 Hz. A homogenous speaker set of male adult speakers of Standard Dutch (as birth, home and work language) was selected, aged between 18 and 50. Speakers conversed with male or female interlocutors, talking about a topic of their choice.

For English, recordings of mock police interviews from the DyViS corpus were used (Nolan et al. 2009). The full corpus contains four tasks (a simulated police interview, a telephone conversation with a mock accomplice, text reading, and sentence reading) and was designed to serve as a research tool for studies into speaker characteristics. It contains speech from 100 male Southern British English speakers, aged between 18 and 25. The interviews are conversations between the speaker acting as suspect and a researcher acting as police officer, and their interaction is guided by on-screen maps providing information about what can and what cannot be shared with the police. For the current study, this first task in the series was preferred over the second task (telephone conversations), which were a debriefing of task 1. The debriefing elicits word repetitions, and in conversational speech, content words are shortened in pronunciation when repeated (Bell et al. 2009). In addition to reduced duration, repetition would potentially also affect formant tracks, and therefore the less-practiced pronunciations of task 1 were preferred. Speech was recorded in a sound-treated room at 44.1 kHz (further details in Nolan et al. 2009).

For both datasets, manual orthographic transcripts were available. Using forced alignment, implemented in Praat (Boersma and Weenink 2018), automatic phonemic transcripts were created. The resulting transcripts were not error-free, but sufficient for subsequent manual segmentation of vowel tokens for inclusion in the analysis (see 2.2).

2.2 Segmentation procedure

Using the automatically-generated phonemic transcripts, instances of the vowel [ei] produced by adult male speakers were located in the audio, and each token was manually assessed for inclusion in the analysis set. Tokens were excluded in the case of (1) misidentifications of [ei] by the automatic phoneme assignment, (2) strong reduction or assimilation, resulting in [ei] not being audible or its phonemic nature altered, (3) background noise or an interfering talker, (4) hesitations or false starts in the token-bearing word, or (5) interfering sounds by the speaker, such as laughter. If necessary, the onset and/or offset locations of the vowel were manually adapted from the automatically determined ones. The Dutch data were checked to include only following consonants that were non-approximants (e.g. stops, nasals, fricatives), as following approximants (/r, w, j/) and coda /l/ may elicit less diphthongal behaviour in the vowel nucleus. In total, 3,072 Dutch tokens were manually segmented from 63 speakers (median of 50 tokens per speaker) and 1,809 English tokens from 61 speakers (median of 29 tokens per speaker).

2.3 Acoustic analysis

Per [ei] token, dynamic representations of the formant trajectories were extracted. As the Dutch data were recorded over a telephone band, only F2 and F3 measurements were included (Künzel 2001). For English, all F1-F3 measurements were initially extracted.

Measurements were taken using Praat (Boersma and Weenink 2018). Starting from the default analysis range proposed by Praat, three formants in 3,0 kHz for male speakers, the latter range was manually adjusted by speaker if this resulted in better fits of the visible formant tracks to the spectrograms. Through the Burg method, formants were measured at nine steps, spaced at equal time intervals throughout the vowel (10–90% of its duration). If all nine measurements per token were available, a cubic fit of these measurements was determined using the *poly*() function in R: $f(t) = a_0 + a_1 \cdot t + a_2 \cdot t^2 + a_3 \cdot t^3$. The cubic fit was preferred over a quadratic fit, because of higher average R^2 values for model fit, and because the transition from /e/ towards /i/ can in principle contain multiple inflection points. If not all subsequent measurements were available, no fit was done. Especially for the English F3 data there were many missing values. Therefore, F3 was excluded as a parameter for English.

For Dutch F2 and F3, 2,314 tokens remained, and for English F1 and F2 1,783 tokens remained. The average R^2 values for model fit were 88–89% for Dutch and 69–72% for English. For each vowel, its duration as well as fundamental frequency (F0, in Hertz) over the mid-50% of the vowel were also measured. The latter was done using an accurate autocorrelation method in Praat, over the 70–350 Hz range.

2.4 Statistical analysis

To evaluate the relative contribution of formant dynamics to the speaker-specificity of the vowel [ei] the strength of evidence was computed as the likelihood ratio (LR) of two conditional probabilities (e.g. Robertson, Vignaux and Berger 2016); the probability of obtaining the evidence under the assumption that the speech samples came from the same speaker, divided by the probability of obtaining the evidence under the assumption that different speech fragments came from different speakers. A MATLAB implementation (Morrison 2007) of the formula proposed by Aitken and Lucy (2004) was used. LRs were computed for same-speaker and different-speaker comparisons. The former ideally yield LRs (well) above one, whereas the latter yield LRs between zero and one. Because it is customary to convert LRs to log-LRs (LLRs), the criterion separating ideal same-speaker versus different speaker scores is placed at zero.

The strength of evidence was computed for different LR systems, in each of the languages. In this study, the different systems only varied in acoustic predictor set (see Morrison 2013, p. 174, for possible definitions of 'system'): (1) formant intercepts only (i.e. the a_0 coefficient from the cubic fits), (2) formant intercepts plus dynamic coefficient (i.e. all coefficients), and (3) all formant coefficients plus vowel duration and mean F0 over the mid-section of the vowel. Remember that for Dutch F2 and F3 were included, whereas for English F1 and F2 were included. Because formant intercepts taken from fits are not the same static formant representation as mid-vowel averaged measurements, a control experiment was set up for Dutch where also an LR system was built including only the mid-formant measurements. The assumption was that both types of static vowel representations, mid-formant measurements and formant intercepts, would contain equal amounts of speaker-dependent information. For F2, the Pearson correlation between formant intercept a_0 and the mid-vowel measurement was $r = .97$ ($p < .001$), for F3 this correlation was $r = .94$ ($p < .001$).

The data per language were each randomly divided into three sets: a training set (used for the computation of calibration parameters), a reference set (background data), and a test set. A third of the available speakers were randomly assigned to each set. For Dutch, each set had 21 speakers, which means that there were 21 same-speaker comparisons and 210 different-speaker comparisons ([21×20]/2). For English, the training set had 21 speakers, whereas the other two had 20. For the training set there were 21 same-speaker comparisons and 210 different-speaker comparisons, whereas for the test set there were 20 same-speaker comparisons and 190 different-speaker comparisons ([20×19]/2). Because there was only one recording

available per speaker, speaker data was divided into first and second halves to allow for same-speaker comparisons. In comparison with speech collections that have multiple recordings per speaker, system performance may be overestimated here (Enzinger and Morrison 2012).

To compute LLRs for the multivariate acoustic representation of /ei/ tokens, first the training set was used to generate same-speaker and different-speaker scores, using the algorithm that models within-speaker variance with a normal distribution, and between-speaker variance using multivariate kernel density (see Aitken and Lucy 2004). Next, these scores were used to compute a logistic regression slope and shift for the purposes of calibration of the test set (Brümmer and du Preez 2006), using a MATLAB implementation by Morrison (2009b). Then, scores were computed for the test set, and calibrated to obtain same-speaker and different-speaker LLRs. Moreover, to capture the precision of the results, the Bayesian equivalent of the 95% confidence interval for the mean, i.e. the 95% credible interval (CI) was computed for the LLRs of the different-speaker comparisons using the non-parametric procedure explained in Morrison et al. (2011). The CI could not be computed for the same-speaker comparisons, because without non-contemporaneous recordings it is not possible to re-run an independent experiment with the same amount of data from the same speakers. Therefore, as an additional test of the precision of the results, the experiment was repeated 10 times, drawing different random samples of training, reference and test sets each time. Different draws vary somewhat in the exact LR results and in performance (Wang et al. 2019). By using a seed for the random generator the same 10 set-combinations were used per LR-system.

The systems differing in acoustic-phonetic predictor sets were evaluated per language through the median LLRs as well as system performance measures. The distance between the median LLR for same-speaker comparisons versus that of different-speaker comparisons is representative of the system's ability to separate the two types of comparisons, and therefore speakers. The first performance measure was the log-likelihood ratio cost function (C_{llr}, Brümmer and du Preez 2006): correct system decisions decrease C_{llr} and more so when LRs are further away from 1, whereas incorrect decisions increase C_{llr} and more so when LRs are further away from 1 (Morrison 2011). It is a measure of the validity of a system. Moreover, C_{llr_min} gives the Cllr when calibration error is minimized (Brümmer and du Preez 2006). Finally, Equal Error Rate was determined, which expresses the percentage of errors where the false-alarm rate (incorrect same-speaker decisions) equals the false rejection rate (incorrect different-speaker decisions). For all performance measures holds that lower values are better, with a minimum at zero, and they were computed using the *sretools* package (Van Leeuwen 2008) in R.

2.5 Results

2.5.1 [ei] in Dutch telephone conversations

The average F2 and F3 trajectories for Dutch [ei] that were fitted using cubic polynomials are shown in Figure 1. F2 increases over time, whereas F3 shows an increase followed by a small decrease. The mean vowel duration was 113 ms (sd = 47 ms) and the mean F0 was 127 Hz (sd = 31 Hz).

[FIGURE 1 NEAR HERE]

Figure 1: F2 and F3 trajectories across Dutch speakers' realizations of [ei], N = 2,314. The 95% confidence intervals of the means is shown in shading.

The performance of each of the three LR-systems using Dutch [ei] parameters is given in Table 1. Remember that the systems differed in parameter sets: 1) formant intercepts only, 2) all

formant coefficients, and 3) all phonetic-acoustic parameters. Each system was run 10 times, using different draws from the dataset, thus allowing for the presentation for medians and ranges of the relevant measures. In addition, results are given for the control condition where not the formant intercepts for F2 and F3, but rather the formant measurements at 50% into the vowel were taken as system input.

Table 1: LR results for four predictor sets, including the main experiment and a control experiment. Results include same-speaker LLRs (LLR_{SS}), different-speaker LLRs (LLR_{DS}), and performance measures Cllr, $Cllr_{min}$ and EER. Per predictor set, the median 95% credible intervals are given for the different-speaker LLRs. Moreover, the median and range [minimum, maximum] are given for each measure.

Predictors		LLR_{SS}	LLR_{DS}	Cllr	$Cllr_{min}$	EER
a_0	median	1.13	−2.95 (±1.98)	0.67	0.44	15.3
	range	[0.80, 1.79]	[−8.04, 0.02]	[0.45, 1.39]	[0.31, 0.60]	[10.5, 19.7]
a_0-a_3	median	1.43	−3.07 (±1.92)	0.61	0.42	14.0
	range	[1.10, 2.33]	[−6.00, −0.03]	[0.40, 1.48]	[0.22, 0.65]	[8.5, 20.9]
all	median	1.76	−4.27 (±1.93)	0.52	0.30	9.7
	range	[1.51, 2.77]	[−6.38, −0.40]	[0.31, 0.95]	[0.17, 0.53]	[6.3, 16.9]
mid formant	median	1.14	−2.98 (±1.69)	0.63	0.47	18.2
	range	[0.85, 1.48]	[−4.60, 0.04]	[0.45, 1.22]	[0.33, 0.64]	[10.5, 19.8]

The results show that median same-speaker LLRs increase as the number of parameters increases, whereas median different-speaker LLRs remain consistent between systems with one or more formant coefficients, and lower for the all-predictor system. The log-likelihood ratio cost, Cllr, lowers when more acoustic parameters are included. This is also the case for $Cllr_{min}$, and for the EER, thus demonstrating improved system performance as more predictors are added. The absolute numbers for each of these measures show that the Dutch vowel [ei] on its own has restricted discriminatory power. This is also reflected in Figure 2, the Tippett plot for the results of the first repetition in each experiment. At the same time, the plot captures the small improvement in performance as a function of LR system; the inclusion of formant parameters beyond the intercept, and also the addition of the further parameters F0 and duration, improve the separation between same-speaker and different-speaker LLRs. Finally, the results of the intercept-system and that of the control system with mid-formant values are comparable.

[FIGURE 2 NEAR HERE]

Figure 2: Tippett plot for repetition 1 data by LR system (DS = different-speaker, SS = same-speaker).

2.5.2 [ei] in English mock police interviews

The average F1 and F2 trajectories for English [ei] that were fitted using cubic polynomials are shown in Figure 2. These reflect the decrease in F1 and the increase in F2 over the course of the vowel. Mean vowel duration was 105 ms (SD = 34 ms) and mean F0 was 108 Hz (SD = 20 Hz).

[FIGURE 3 NEAR HERE]

Figure 3: F1 and F2 trajectories across English speakers' realizations of [ei], N = 1,785. The 95% confidence intervals of the means is shown in shading.

Table 2 shows that the same-speaker LLRs become more positive as more acoustic parameters are added. Different-speaker LLRs become weaker from the system without to the system with formant dynamics, but stronger again when all parameters are included. Cllr lowers from the system without to that with dynamics, and further when all parameters are included. $Cllr_{min}$ remains comparable between the intercept and all-coefficients systems, but lowers for the all-parameter system. Just as Cllr, EER lowers when more acoustic parameters are included. The addition of formant dynamics has a relatively small impact on separating same-speaker from different-speaker LLRs, when compared with the further addition of non-formant information (F0 and duration). System performance measures, however, show improvement with the addition of dynamic formant information; there again is a larger change when F0 and duration are included.

Figure 4 shows the Tippett plot for results from repetition 1. It reflects that the addition of dynamic formant information in that particular case did not separate different-speaker scores further from same-speaker scores. Best performance, just as across repetitions, is found when all acoustic parameters are included.

Table 2: LR results for three predictor sets, including same-speaker LLRs (LLR_{SS}), different-speaker LLRs (LLR_{DS}), and performance measures Cllr, $Cllr_{min}$ and EER. Per predictor set, the results include the median 95% credible interval for the different-speaker LLRs. Moreover, the median and range [minimum, maximum] are given for each measure.

Predictors		LLR_{SS}	LLR_{DS}	Cllr	$Cllr_{min}$	EER
a_0	median	1.70	−6.02 (±2.39)	0.48	0.30	15.5
	range	[1.48, 2.11]	[−7.32, −2.93]	[0.38, 0.62]	[0.32, 0.49]	[10.9, 18.6]
a_0-a_3	median	2.03	−5.61 (±2.29)	0.38	0.30	10.4
	range	[1.63, 2.84]	[−11.85, −2.10]	[0.27, 0.74]	[0.18, 0.42]	[5.6, 14.9]
all	median	3.36	−10.08 (±3.03)	0.28	0.16	5.2
	range	[2.91, 5.79]	[−46.07, −6.64]	[0.19, 0.62]	[0.13, 0.33]	[4.1, 9.4]

[FIGURE 4 NEAR HERE]

Figure 4: Tippett plot for repetition 1 data by LR system (DS = different-speaker, SS = same-speaker).

2.6 Intermediate discussion

This experiment assessed the contributions of formant dynamics and formant intercepts to speaker-dependent information in conversational speech. In separate experiments, [ei] productions from Dutch and from English speakers were evaluated. It was expected that in conversational speech dynamic formant information would add speaker-dependent information to static formant information in diphthongal vowels. The results from these experiments confirmed this prediction. Even in the case where the separation between same-speaker and different-speaker scores benefited little from the addition of dynamic information, system performance measures still improved. Moreover, additional acoustic parameters improved performance further.

3 Experiment 2: The relative contribution of formant dynamics as a function of vowel quality

The question addressed in this experiment was whether the contribution of dynamic formant information to speaker-dependency would be larger in a diphthongal versus monophthongal vowel. It was expected that the former type of vowel would benefit most from the inclusion of higher cubic fit coefficients in an LR system. Because different speech sounds inherently vary in speaker-dependent information (e.g. Van den Heuvel 1996), not the absolute results but

rather the relative contribution made by the formant coefficients to the formant intercept baseline is relevant for answering this question.

3.1 Method and materials

The speaker-specific information contributed by the monophthong /a:/ and the phonetic diphthong /e:/ were compared. The vowels came from the same Dutch data as mentioned in section 2.1. For the same 48 male speakers, data from both vowels were available (/a:/: N = 2,571 tokens; /e:/: N = 1,836 tokens). The segmentation procedure was explained above in 2.2.

Even though F1 for /a:/ lies well above the lower cut-off frequency for telephone speech, only F2 and F3 were included in this analysis, for comparability with the /e:/ data by the same speakers. Acoustic formant measurements were performed and fitted with cubic polynomials as explained in section 2.3. For the current experiment, F0 and duration were not included.

3.2 Statistical analysis

Because there were fewer speakers in this experiment than in either dataset of experiment 1, LRs were determined using a leave-one-out method. Hughes (2017) showed that when working with separate training, reference and test sets, the minimum set size is about 20 speakers and especially for the training set. In the current experiment there were 48 same-speaker comparisons and 1,128 (= [48×47]/2) different-speaker comparisons for each vowel. Because there was only one recording available per speaker, speaker data was divided into first and second halves to allow for same-speaker comparisons. In same-speaker comparisons, a speaker's first half was compared to their second half. In different-speaker comparisons, a speaker's first half was compared to a higher-numbered speaker's second half.

The performance of two LR systems was evaluated for both vowels separately. The first system included the formant intercepts only, and the second system used all formant coefficients. To compute the LLRs, the first step of score calculation was a sequential leave-one-out implementation of the method explained in section 2.4. To compute the score per same- or different-speaker pair, the reference set consisted of all speaker but the one(s) to be compared. Calibration was also performed using leave-one-out cross-validation, in which again the speaker or speakers from whom a score was calibrated were left out of the data set to determine the logistic regression coefficients for score-to-LR transformation.

Performance of the different systems on the two vowels was evaluated in the same way as in Experiment 1 (see section 2.4). Finally, the 95% CI was computed for the different-speaker comparisons using the non-parametric procedure from Morrison et al. (2011).

3.3 Results and discussion

In Figure 5 the cross-speaker F2 and F3 formant tracks are given for both vowels. As can be seen, F2 of [e:] shows more change than that of [a:].

[FIGURE 5 NEAR HERE]

Figure 5: F2 and F3 trajectories across Dutch speakers' realizations of /a:/ (solid) and /e:/ (dotted). The 95% confidence intervals of the means is shown in shading.

LR results are presented in Table 3. For the phonologically monophthongal vowel /a:/ there was no positive effect on system performance as a function of adding the dynamic coefficients to the static formant intercepts; the median same-speaker LLR stays at the same level, and the median different-speaker LLR weakens (rather than strengthens) slightly. The performance measures Cllr, $Cllr_{min}$ and EER do not improve when dynamic coefficients are added.

For the diphthongal vowel [ei], results change when dynamic coefficients are added to the formant intercepts; the separation between median LLRs for same- and different-speaker comparisons becomes larger, and performance measures lower, thus reflecting less errors and higher validity (see also Figure 6).

Table 3: LR results for both vowels from two predictor sets each, including same-speaker LLRs (LLR_{SS}), different-speaker LLRs (LLR_{DS}), and performance measures Cllr, $Cllr_{min}$ and EER. The 95% credible interval for the different speaker LLRs is also given.

Vowel	Predictor set	LLR_{SS}	LLR_{DS}	Cllr	$Cllr_{min}$	EER
/a:/	Formant intercepts	1.28	−2.06 (±1.67)	0.61	0.56	21.2
	All formant coefficients	1.27	−1.74 (±1.44)	0.63	0.58	22.4
/e:/	Formant intercepts	1.13	−1.37 (±1.53)	0.74	0.53	16.4
	All formant coefficients	1.50	−2.08 (±1.65)	0.63	0.45	14.0

[FIGURE 6 NEAR HERE]

Figure 6: Tippett plots for the [ei] and [a:] vowel systems, based on F2 and F3 intercepts (dashed) or all coefficients (solid), (DS = different-speaker, SS = same-speaker).

4. General discussion

The first question in this investigation was whether in conversational speech dynamic formant coefficients in a diphthongal vowel contribute speaker-dependent information over that provided by static formant representations. This question was tested on two independent databases, from two languages, where the similarly realized [ei] was a phonetic diphthong in Dutch and a phonological one in English. The second question was whether the speaker-specificity of dynamic formant information in conversational speech varies with the inherent variation in a vowel. To answer this question, the speaker-dependent information in the monophthong [a:] versus diphthong [ei] was compared, with vowels produced by the same speakers.

An advantage of dynamic formant over static formant representations was found for the inherently dynamic vowel [ei] analysed in conversational speech. This was the case for both languages, Dutch and English, in spite of ample variation in segmental contexts and linguistic positions that tokens were sampled from (various syllable structures, stressed and unstressed syllables, function and content words). This suggests that also in the variable contexts found in recordings used for forensic speaker comparisons, the dynamic component in formant representations is likely to add speaker-specific information. When comparing the strength-of-evidence and performance obtained in Experiment 1 to results from earlier studies, both were lower than in more controlled speech (Morrison 2009a), but relatively consistent with other results on vowels sampled from the DyViS database's conversational speech (Gold 2014, ch. 5; Hughes et al. 2016). Note, however, that in the current study contemporaneous recordings were used, which may have led to an over-estimation of performance (Enzinger and Morrison, 2012). In spite of this, the goal of this investigation was not to arrive at optimal system performance, but rather to investigate the relative contributions of static and dynamic information in less and more inherently-variable vowels.

When comparing performance between the two diphthong experiments, there also were some differences. F2 and F3 were included for Dutch, whereas F1 and F2 were used for English. This difference was due to differences in the recording conditions between the two speech collections. When comparing the results from the LR systems containing all formant coefficients of [ei], but not the additional acoustic parameters, the F1-F2 system here performed better than the F2-F3 system in terms of absolute LLRs, and of performance measures; in terms

of the added value of the dynamics, systems performed comparably. A direct comparison of particular formant combinations based on results from the current study, however, is confounded by these combinations (F1-F2 versus F2-F3) being extracted from different speech collections (face-to-face and telephone conversation) in different languages, recorded in different ways. The Dutch collection, for instance, was recorded under less-controlled circumstances in speakers' homes rather than in a studio, thus potentially affecting the quality of measurements. Moreover, the Dutch speaking task was much less-controlled than that in English, thus potentially affecting sources of within- and between-speaker variation. Also, even though speakers were selected to speak a standard variety of their language, the larger age range of the Dutch speakers may have influenced between-speaker variation. These differences may explain why the present comparison of formant combinations seems to run counter to comparisons reported in the literature. In contrast to the current study, in both Morrison (2009a) and Hughes et al. (2016) the performance of different formant combinations on inherently-dynamic vowels was compared within one data set. Morrison (2009a) included several diphthongs from read speech, and for Australian English /eɪ/ in particular found the best performance for the F2-F3 combination (also better than F1-F3). When comparing performance measures of quadratic fits on *um* between F1-F2 and F2-F3 systems (see Hughes et al. 2016, Fig. 10), the latter system, i.e. F2-F3, had C_{llr}s and EERs closer to zero. A slight advantage of an F2-F3 relative to an F1-F2 system was also found for long-term formants by Gold (2014, ch. 5). Across studies that included direct comparisons of formant combinations using the same data, F2-F3 representations thus seem to show better speaker-specificity results. This would be a fortunate situation for casework that includes telephone speech, where F1 often cannot be used (Künzel 2001).

The additional acoustic parameters used in Experiment 1, F0 and duration, improved performance over that of the formants only. This finding, in itself, is not surprising, but what is interesting is that the improvement was relatively large in English, and also larger than in Dutch. This difference may be attributed to the more controlled circumstances under which the English collection was recorded, and maybe also the larger variation in speaker characteristics in the Dutch collection. This seems reflected in an inspection of by-speaker variances for F0, which on average varied much less in English than in Dutch. This result furthermore suggests that the added benefit of F0 for speaker-specificity is much smaller in spontaneous conversational speech, as in the Dutch collection, relative to task-guided conversation, as in the English collection. This is consistent with reports on the usefulness of F0 for forensic speaker comparisons (e.g. Gold and French, 2011), which is deemed relatively low.

When comparing speech features contributing to the evidential value of monophthongal /aː/ versus diphthongal /eː/ in Experiment 2, it was found that only in the latter case there was a benefit from the addition of fit coefficients beyond the intercept, that is of formant dynamics. This result on the monophthongal vowel is consistent with results reported on the vowel in hesitation marker *uh* (Hughes et al., 2016), but differs from results on monophthongal vowels reported in Rose (2015) and Feljovà et al. (2013). In the latter two studies an added benefit of dynamic vowel representations was found. In its design, the current study was most similar to Hughes et al. (2016), who sampled tokens from conversational speech, and also used the LR framework to assess speaker-specificity. Rose (2015) also computed LRs, but used tokens from only eight different word contexts. This restriction on the amount of contextual variation may contribute to explaining the difference; less contextual variation is likely to reduce the amount of within-speaker variance. Feljovà et al. (2013) used read speech in combination with linear discriminant analysis. The difference in contextual and therefore articulatory control from the current study, in addition to a difference in speaker-dependency measure, may here have contributed to the varying results. Taking the results from the different studies together, the added benefit of dynamic formant information seems to be strongly reduced –in spite of vowel-

inherent spectral change even in monophthongs– under conditions of contextual and linguistic variation. In the current study, the relatively speaker-specific Dutch vowel [aː] was sampled from a range of different words, in various utterance positions, thus leaving no speaker-specific information in its dynamics. When sampled from a constant context, such as the central vowel in hesitation marker *um*, monophthongs in conversational speech may still benefit from dynamic representations (Hughes et al. 2016).

There are some limitations to the current investigation. Even though Experiments 1a and 2 were done using telephone speech, the lack of non-contemporaneous recordings means that the results may overestimate performance. In acoustic-phonetic forensic speaker comparisons that would include formant analysis, however, other speaker characteristics would be included as well (e.g. other speech sounds, F0, tempo, voice quality). Moreover, only one diphthongal vowel from conversational speech was examined (but in two languages). The choice for this vowel may in fact provide a slight overestimation of diphthongal vowels' information content in general, as in Morrison (2009a) the /ɐɪ/ vowel was found to perform best out of five Australian English diphthongs. Moreover, even though males are much more frequently encountered in casework than females, results cannot necessarily be generalized across the sexes. Not only are females' formants generally higher, female articulation tends to be less reduced (Byrd 1994), thus potentially affecting within- and between-speaker variation. Future work intends to address mainly the issues of including forensically-realistic recordings and using non-contemporaneous speech data.

5. Conclusion

Results showed that in an inherently-dynamic vowel the formant dynamics, as quantified by higher polynomial fit coefficients, contribute speaker-dependent information to static formant information, as quantified by the polynomial fit intercept. Because this result is based on two independent datasets of conversational speech, which also vary in their design and composition, it is expected to generalize to other inherently-dynamic vowels in different languages. A second experiment, comparing diphthong to monophthong realizations, showed that in the monophthong, dynamic formant information did not add to static formant information. For casework, these results imply that dynamic formant representations, such as polynomial fits, may be unnecessarily complex for monophthongs. For diphthongs, and given earlier work also monophthongs in constant phonetic contexts, dynamic formant representations contain most speaker-specific information.

Acknowledgements

This research was supported by a VIDI grant (276-75-010) from The Netherlands Organisation for Scientific research. I would like to thank two anonymous reviewers for their constructive feedback.

References

Adank. P., van Hout, R. and Van de Velde, H. (2007) An acoustic description of northern and southern standard Dutch II: Regional varieties. *The Journal of the Acoustical Society of America* 121(2): 1130–1141.

Bell, A., Brenier, J. M., Gregory, M., Girand, C. and Jurafsky, D. (2009) Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60: 92–111.

Blumstein, S. E. and Stevens, K. N. (1981) Phonetic features and acoustic invariance in speech. *Cognition* 10: 25–32.

Boersma, P. and Weenink, D. (2018) Praat. Doing phonetics by computer [Computer program]. Version 6.0.42.

Byrd, D. (1994) Relations of sex and dialect to reduction. *Speech Communication* 15(1-2): 39–54.

Brümmer, N. and Du Preez, J. (2006) Application-independent evaluation of speaker detection. *Computer Speech & Language* 20(2-3): 230–275.

Enzinger, E. and Morrison, G. S. (2012) The importance of using between-session test data in evaluating the performance of forensic-voice-comparison systems. *In Proceedings of the 14th Australasian International Conference on Speech Science and Technology* 137–140, Sydney, Australia.

Enzinger, E. and Morrison, G. S. (2017) Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case. *Forensic Science International* 277: 30-40.

Fejlovà, D., Lukeš, D. and Skarnitzl, R. (2013) Formant contours in Czech vowels: Speaker-discriminating potential. *Proceedings of Interspeech 2013* 3182–3186, 25–29 August 2013, Lyon, France.

Franco-Pedroso, J. and Gonzalez-Rodriguez, J. (2016) Linguistically-constrained formant-based i-vectors for automatic speaker recognition. *Speech Communication* 76: 61–81.

Gold, E. A. (2014) *Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters*. PhD dissertation, University of York, UK.

Gussenhoven, C. (1999) Dutch. In International Phonetic Association, and International Phonetic Association Staff (ed.) *Handbook of the International Phonetic Association. A guide to the use of the International Phonetic Alphabet* 74–77, Cambridge: Cambridge University Press.

Hughes, V., Wood, S. and Foulkes, P. (2016) Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech, Language and the Law* 23(1): 99–132.

Hughes, V. (2017) Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough? *Speech Communication* 94: 15–29.

Ingram, J. C. L., Prandolini, R. and Ong, S. (1996) Formant trajectories as indices of phonetic variation for speaker identification. *Forensic Linguistics* 3(1): 129–145.

Johnson, K., Ladefoged, P. and Lindau, M. (1993) Individual differences in vowel production. *The Journal of the Acoustical Society of America* 94(2): 701–714.

Jones, D. (1957) *An outline of English phonetics*. Cambridge: W. Heffer and Sons Ltd.

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.

Aitken, C. G. G. and Lucy, D. (2004) Evaluation of trace evidence in the form of multivariate data. *Applied Statistics* 53: 109–122.

McDougall, K. (2004) Speaker-specific formant dynamics: an experiment on Australian English /aɪ/. *International Journal of Speech, Language and the Law* 11(1): 103–130.

McDougall, K. (2006) Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies. *International Journal of Speech, Language and the Law* 13(1): 89–126.

McDougall, K. and Nolan, F. (2007) Discrimination of Speakers Using the Formant Dynamics of /u:/ in British English In J. Trouvain and W. Barry (eds) *Proceedings of the 16th International Congress of Phonetic Sciences* 1825–1828, 6–10 August 2007, Saarbrücken, Germany.

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

Morrison, G. S. and Nearey, T. M. (2007) Testing theories of vowel inherent spectral change. *Journal of the Acoustical Society of America* 122: EL15–22.

Morrison, G. S. (2007) Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation, Downloaded from https://geoff-morrison.net/#MVKD, last visited on 28-11-2019.

Morrison, G. S. (2008) Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aɪ/. *International Journal of Speech, Language and the Law* 15(2): 249–266.

Morrison, G. S. (2009a) Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *Journal of the Acoustical Society of America* 125(4): 2387–2397.

Morrison, G. S. (2009b) train_llr_fusion_robust.m, Downloaded from https://geoff-morrison.net/#TrainFus, last visited on 28-11-2019.

Morrison, G. S. (2011) A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM–UBM). *Speech Communication* 53: 242–256.

Morrison, G. S. (2013) Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences* 45(2): 173–197.

Morrison, G. S., Zhang, C. and Rose, P. (2011) An empirical estimate of the precision of likelihood ratios form a forensic-voice-comparison system. *Forensic Science International* 208: 59–65.

Nearey, T. M. and Assman, P. F. (1986) Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America* 80: 1297–1308.

Nolan, F. and Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law* 12(2): 143–173.

Nolan, F., McDougall, K., de Jong, G. and Hudson, T. (2009) The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law* 16(1): 31–57.

Oostdijk, N. H. J. (2000) Het Corpus Gesproken Nederlands [The Spoken Dutch corpus]. *Nederlandse Taalkunde* 5: 280–284.

Peterson, G. E. and Barney, H. L. (1952) Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America* 24(2): 175–184.

Roach, P. (2004) British English. Received pronunciation. *Journal of the International Phonetic Association* 34(2): 239–245.

Robertson, B., Vignaux, G. A. and Berger, C. E. (2016) *Interpreting evidence: evaluating forensic science in the courtroom*. Chichester: John Wiley & Sons.

Rose, P. (1999) Long- and short-term within-speaker differences in the formants of Australian *hello*. *Journal of the International Phonetic Association* 29(1): 1–31.

Rose, P. (2002) *Forensic speaker identification*. London and New York: Taylor & Francis.

Rose, P. (2015) Forensic voice comparison with monophthongal formant trajectories-a likelihood ratio-based discrimination of "schwa" vowel acoustics in a close social group of young Australian females. *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing* 4819–4823.

Schindler, C. and Draxler, C. (2013) Using spectral moments as a speaker specific feature in nasals and fricatives. *Proceedings of Interspeech* 2793–2796, Lyon, France, 25–29 August 2013.

Thaitechawat, S. and Foulkes, P. (2011) Discrimination of speakers using tone and formant dynamics in Thai. *Proceedings of ICPhS XVII* 1978–1981, Hong Kong, 17–21 August 2011.

Van den Heuvel, H. (1996) *Speaker variability in acoustic properties of Dutch phoneme realisations*. PhD dissertation, Radboud University Nijmegen.

Van de Velde, H. (1996) *Variatie en verandering in het gesproken Standaard-Nederlands*. Nijmegen: Katholieke Universiteit Nijmegen

Van Leeuwen, D. A. (2008) SRE-tools, a software package for calculating performance metrics for NIST speaker recognition evaluations. Downloaded from http://sretools.googlepages.com/, last visited on 02-03-2020.

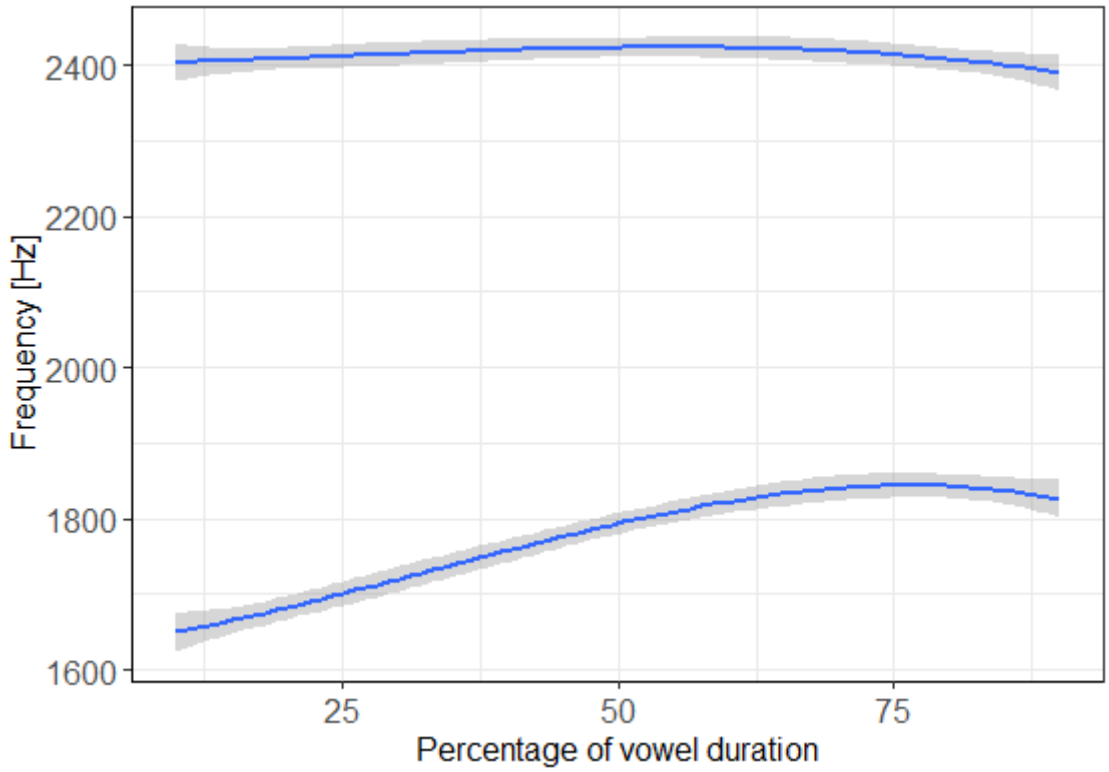
Voeten, C. C. (submitted) *The adoption of sound change. Synchronic and diachronic processing of regional variation in Dutch*. PhD dissertation, Leiden University.

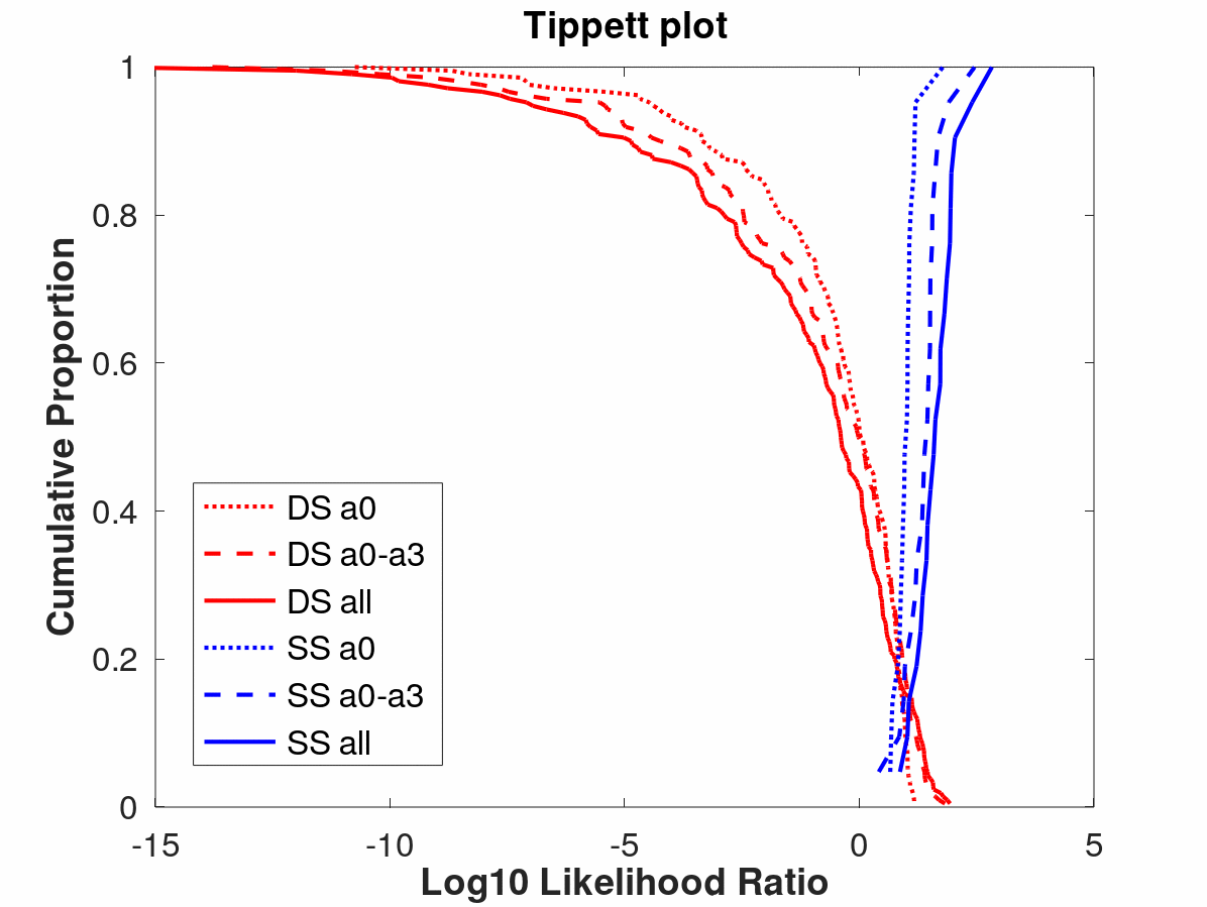
Wang, B. X., Hughes, V. and Foulkes, P. (2019) The effect of speaker sampling in likelihood ratio based forensic voice comparison. *International Journal of Speech Language and the Law* 26(1): 97–120.

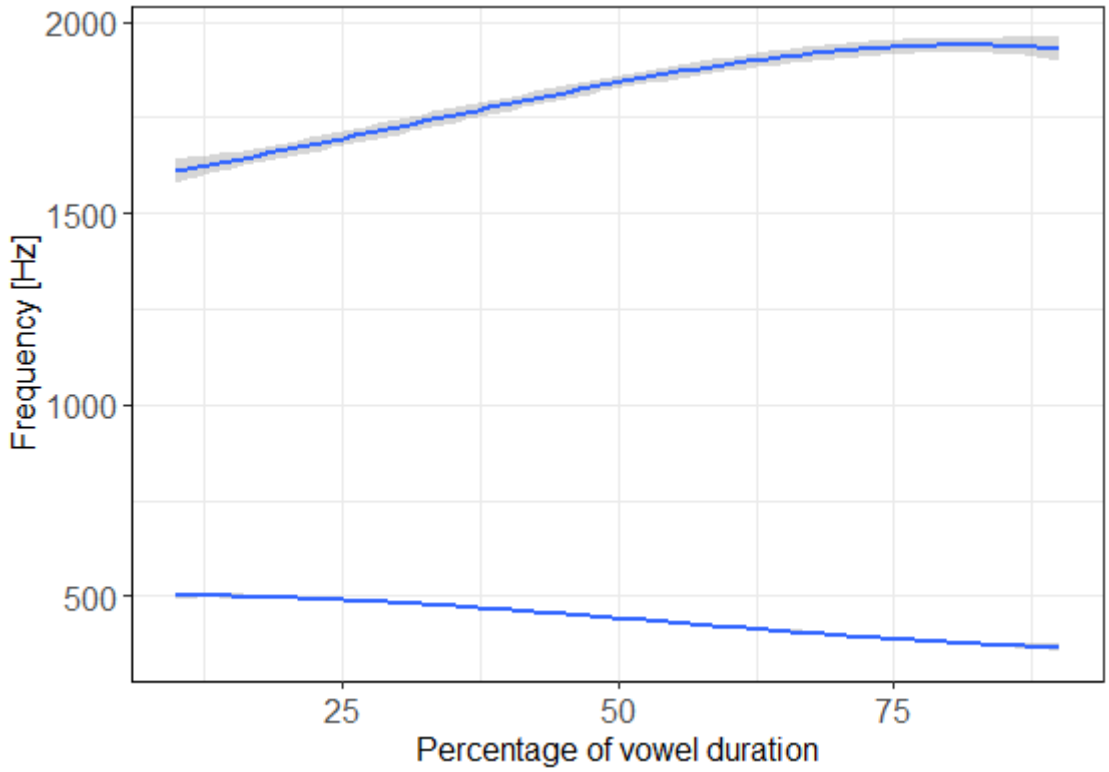
Weirich, M. and Simpson, A. P. (2018) Individual differences in acoustic and articulatory undershoot in a German diphthong – Variation between male and female speakers. *Journal of Phonetics* 71: 35–50.

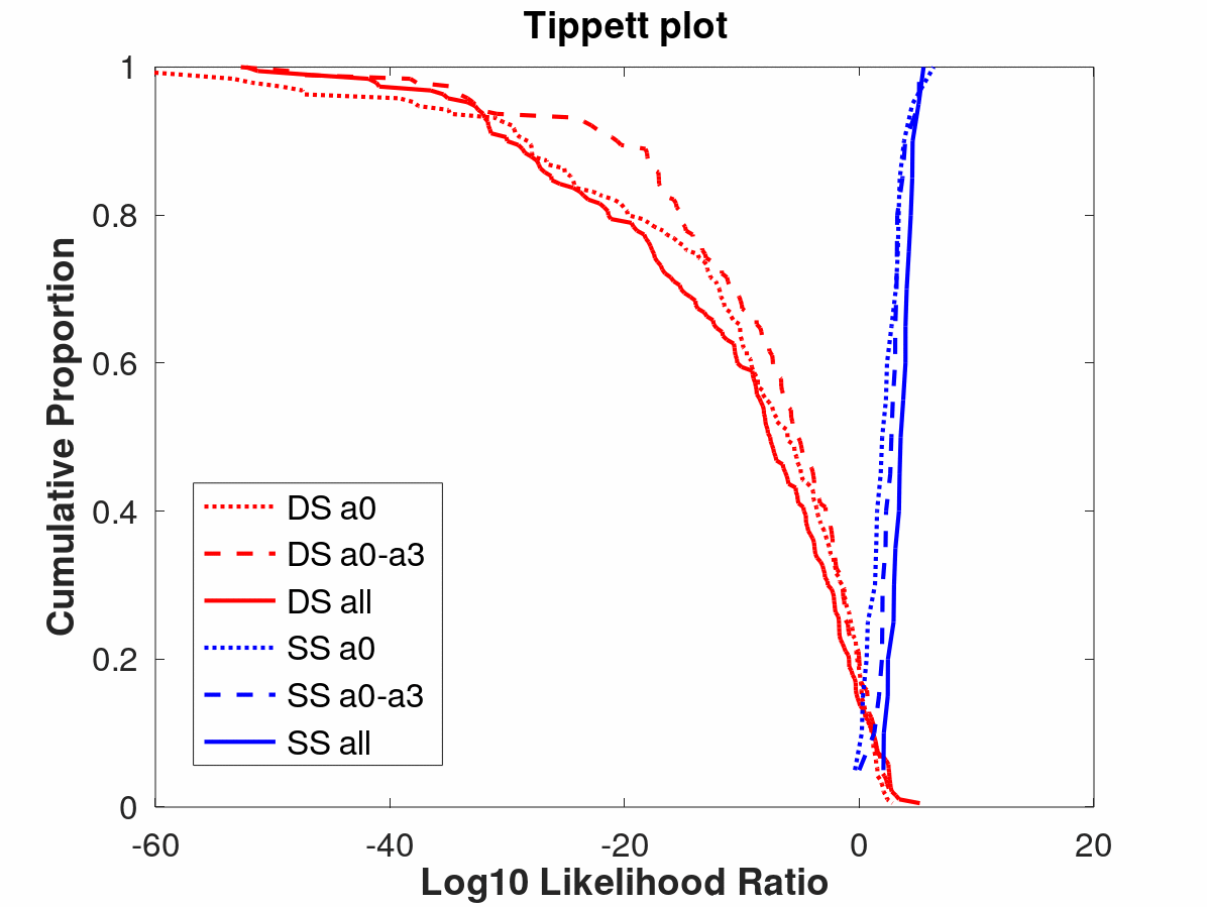
Zhang, C., Morrison, G. S., Enzinger, E. and Ochoa, F. (2013) Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison–female voices. *Speech Communication* 55(6): 796–813.

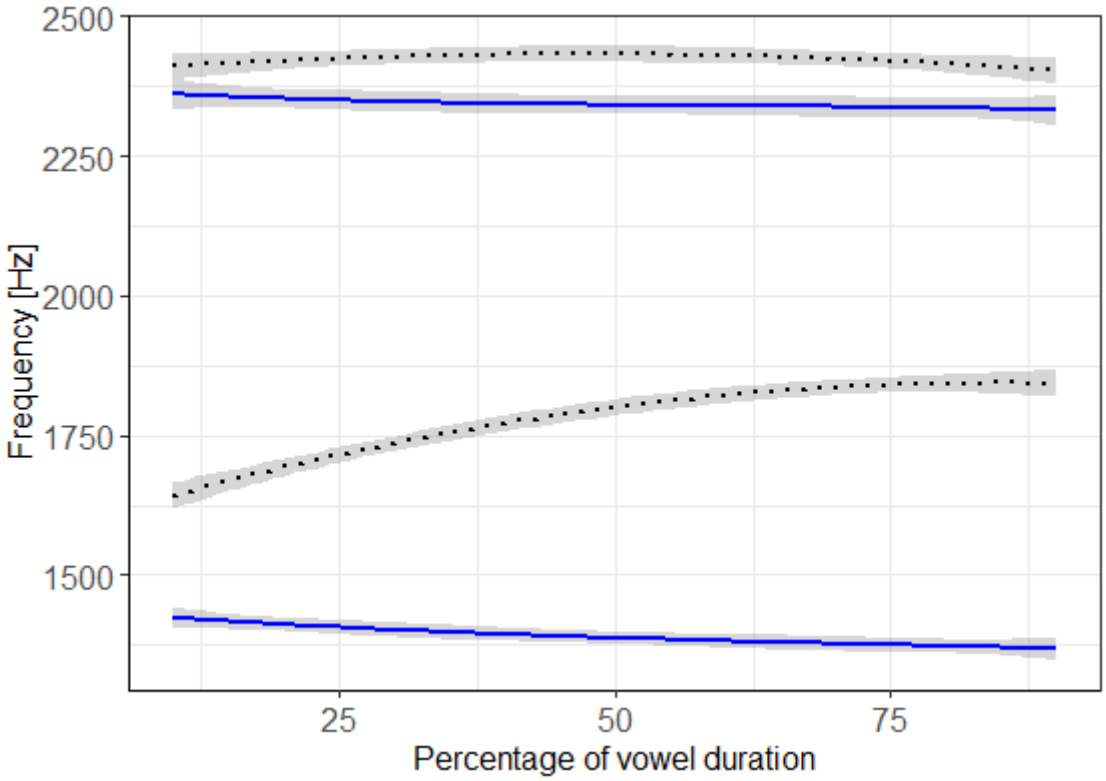
Zuo, D. and Mok, P. P. K. (2015) Formant dynamics of bilingual identical twins. *Journal of Phonetics* 52: 1–12.

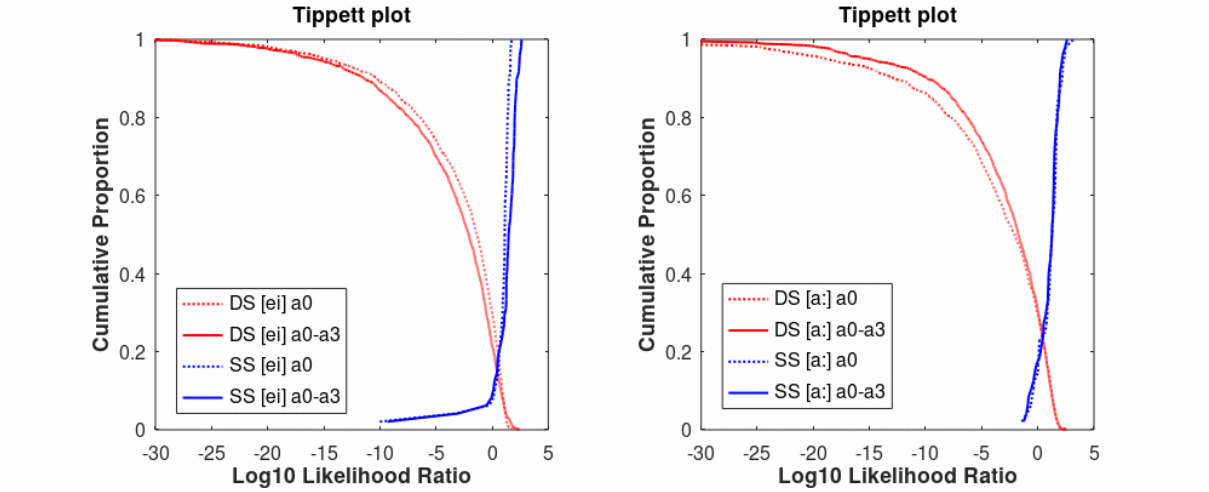

2400
2200
2000
1800
1600
Frequency [Hz]
25
50
75
Percentage of vowel duration


Tippett plot
Cumulative Proportion
Log10 Likelihood Ratio
1
0.8
0.6
0.4
0.2
0
-15
-10
-5
0
5
DS a0
DS a0-a3
DS all
SS a0
SS a0-a3
SS all


2000
1500
1000
500
Frequency [Hz]
25
50
75
Percentage of vowel duration


Tippett plot
Cumulative Proportion
Log10 Likelihood Ratio
1
0.8
0.6
0.4
0.2
0
-60
-40
-20
0
20
DS a0
DS a0-a3
DS all
SS a0
SS a0-a3
SS all


2500
2250
2000
1750
1500
Frequency [Hz]
25
50
75
Percentage of vowel duration


Tippett plot
Cumulative Proportion
Log10 Likelihood Ratio
DS [ei] a0
DS [ei] a0-a3
SS [ei] a0
SS [ei] a0-a3
Tippett plot
Cumulative Proportion
Log10 Likelihood Ratio
DS [a:] a0
DS [a:] a0-a3
SS [a:] a0
SS [a:] a0-a3