The Effect of Word Class on Speaker-dependent Information in the Standard Dutch Vowel /aː/

Willemijn F. L. Heeren

w.f.l.heeren@hum.leidenuniv.nl

Leiden University Centre for Linguistics

Leiden University

Reuvensplaats 3-4

2311 BE Leiden, The Netherlands

Running title: Word Class Effects on Speaker-dependent Information

This is the author submitted version. The published version of the article can be found at https://doi.org/10.1121/10.0002173. To cite, please use:

Heeren, W. F. L. (2020). The Effect of Word Class on Speaker-dependent Information in the Standard Dutch Vowel /aː/. Journal of the Acoustical Society of America, 148(4), 2028-2039.

Abstract

Linguistic structure co-determines how a speech sound is produced. This study therefore investigated whether the speaker-dependent information in the vowel [aː] varies when uttered in different word classes. From two spontaneous speech corpora, [aː] tokens were sampled and annotated for word class (content, function word). This was done for 50 male adult speakers of Standard Dutch in face-to-face speech (N = 3,128 tokens), and another 50 male adult speakers in telephone speech (N = 3,136 tokens). First, the effect of word class on various acoustic variables in spontaneous speech was tested. Results showed that [aː]s were shorter and more centralized in function than content words. Next, tokens were used to assess their speaker-dependent information as a function of word class, by using acoustic-phonetic variables to (a) build speaker classification models, and (b) compute the strength-of-evidence, a technique from forensic phonetics. Speaker-classification performance was somewhat better for content than function words, whereas forensic strength-of-evidence was comparable between the word classes. This seems explained by how these methods weigh between- and within-speaker variation. Because these two sources of variation co-varied in size with word class, acoustic word-class variation is not expected to affect the sampling of tokens in forensic speaker comparisons.

Keywords: speech production (43.70.-h); forensic acoustics (43.72.Uv);

The Effect of Word Class on Speaker-dependent Information in the Standard Dutch Vowel /aː/

I. INTRODUCTION

Speech can be defined as a message carried by a speaker's voice. Speech perception research has provided much evidence that speaker information interacts with the interpretation and memory of a spoken message (e.g., Palmeri *et al.*, 1993; Van Berkum *et al*., 2008). For voice perception, both within- and between-speaker acoustic variation are important (Lavan *et al.*, 2018), whereas the speech production literature shows that speech acoustics, and its variation, depend on linguistic context (e.g., Smorenburg & Heeren, 2020). Taken together, this suggests that also speaker-dependent voice characteristics may be conditioned by linguistic context. Knowledge on how language and voice interact in speech production, however, lags behind; it is the core question of the current paper.

Recent research on voice modelling has investigated which acoustic dimensions may be important for modelling a multi-variate acoustic voice space (see Lee *et al.*, 2019, and references therein), but to the author's knowledge, such research has hardly differentiated between linguistic contexts. There is evidence, however, that speaker-dependent information in an utterance is affected by speech style (e.g., Moos, 2010; Dellwo *et al*., 2015) or speech sound (Van den Heuvel, 1996; Andics, 2013; Kavanagh, 2014). Moreover, Smorenburg and Heeren (2020) recently found that the speaker information contained by the Dutch fricatives /s/ and /x/ to some extent depends on whether the fricative was produced in onset versus coda position. This finding was explained as articulatorily less-demanding positions, such as codas, allowing for more between-speaker variation (see He and Dellwo, 2017). In the present study, the distribution of speaker-dependent information within an utterance is investigated further by examining if differences in vowel pronunciation as a function of word class affect the available speaker information.

In addition to potentially informing voice modelling, the present study is relevant for forensic phonetics, a subfield of phonetics concerned with speaker correlates rather than linguistic ones. A main question is how voices can be characterized acoustically. The outcome of such research feeds into practice; in a forensic speaker comparison (FSC), one or more disputed speech recordings are compared with one or more reference recordings of a suspect in order to investigate whether the recordings might have been produced by the same or by different speakers. To make these comparisons, several methods are in use across the world, varying from auditory examination to acoustic-phonetic measurement to automatic speaker recognition (Morrison *et al.*, 2016; Gold and French, 2019). It is theoretically important to not only compare the disputed and suspect samples to each other, and to thus assess their similarity, but to evaluate the likelihood of this similarity against background population information, to thus assess the typicality of the features under study. Automatic speaker recognition (ASR) by default uses background information, and has the advantages of objectivity and replicability. Even though this method has demonstrated superior performance in telephone-to-telephone speech comparisons (e.g. Zhang *et al.* 2013), it often cannot be applied to case data due to restrictions imposed by data quantity and quality or because ASR is not admissible in the jurisdiction. Moreover, not all types of speech features, such as word use, can be included in ASR. Currently, in international surveys amongst respondents carrying out FSC the majority used an auditory/acoustic-phonetic approach (Morrison *et al.*, 2016; Gold and French, 2019): acoustic-phonetic features are measured in the different speech samples, and used to assess how similar these features are between the suspect and the disputed speaker, relative to how typical they are of speakers in general. Little is known, however, about how the speaker information carried by acoustic-phonetic features depends on the linguistic context from which it is sampled.

Theoretically, discriminative (or: speaker-specific) features exhibit small within-speaker variation while also showing large between-speaker variation, thus differentiating speakers along some feature dimension. Moreover, features which are frequently available in shorter samples and measurable in the low-quality and/or noisy recordings typical of FSC are preferred. In the search for optimal features for acoustic-phonetic FSC, earlier work has compared speaker information carried by different segments (e.g., Van den Heuvel, 1996; Andics, 2013) and different speech styles (Moos, 2010; Dellwo *et al.,* 2015). What is largely lacking from the existing literature, with the exception of Smorenburg and Heeren (2020), is a systematic investigation of how the speaker information carried by a segment may be affected by its position in the utterance *within* the same speech style. A speech sound's acoustics are altered by linguistic structure, such as whether it is realized in a lexically-stressed or a focused position. Therefore, when it – practically – comes to sampling speaker-dependent features optimally for FSC or – theoretically – comes to understanding how voice information is encoded in speech and processed by listeners, it is important to know the distribution of speaker information across an utterance.

A. The interaction between linguistic and speaker-dependent information

Earlier research has shown that vowels tend to carry more speaker-dependent information than consonants, both in production (Van den Heuvel, 1996, p. 145-146) and perception (Andics, 2013, ch. 2). Within the classes of consonants and vowels, there is also variation in speaker-dependent information. Using Dutch CVC words as stimuli, Andics (2013, ch. 2) found that the perceptual discriminability of voices depended on their segmental composition; better results were found for onset /m/ than /l/, nucleus /ɛ/ than /ɔ/, and coda /s/ than /t/. The higher speaker-dependency of /m/ and /s/ was also reported for English read speech by Kavanagh (2014, pp. 387-388), relative to nasals /n/ and /ŋ/, and liquid /l/. Using Dutch read nonsense words, Van den Heuvel (1996) reported similar segmental differences, but he found /n/ to be more speaker-dependent than /m/. A comparison of the three Dutch corner vowels showed that /aː/, which is also used in the present study, contained most speaker-specific information in both the durational and spectral domains, relative to /i/ and /u/ (Van den Heuvel, 1996). An explanation for these differences is mainly given by articulatory differences between speech sounds, also in relation to their neighboring sounds (Smorenburg and Heeren, 2020), together with the anatomical/physiological differences between individual speakers.

Speech sounds differ in how many and which articulators are involved in their production; this creates diversity between speech sounds in the types and amounts of acoustic speaker correlates. An obvious distinction is that between voiced and unvoiced sounds, which relates to involvement of the vocal folds and thus the presence or absence of F0 and its harmonics as a speaker correlate (see Lee *et al.*, 2019). Furthermore, between speakers there are differences in the shapes of the passive articulators (including the teeth, the alveolar ridge and the palate), in the movements of the active articulators (e.g. lips, tongue, vocal folds), and in default articulatory settings (see Laver, 1980, ch. 2). These differences yield speaker-dependent acoustics, an illustration of which can be found in the well-known vowel chart of Peterson and Barney (1952): each of the 76 different speakers produced different combinations of first-second formant values for the same set of vowels. As for the relative contributions of source versus filter variables, Bachorowski and Owren (1999) found that within the same sex, speaker information in vowels was mostly carried by acoustic variables determined by the vocal tract rather than the vocal folds. A possible explanation is that for the majority of same-sex speakers, the within-speaker variation in F0 is relatively large, whereas between-speaker variation in F0 is relatively small.

Different speech styles also cause variation in a speaker's acoustics. In read as opposed to spontaneous German speech, the same speakers produced higher values for their long-term second and third formants (Moos, 2010). Additional acoustic variables cueing read versus spontaneous speech to listeners were reported by Laan (1997); Dutch read speech tended to be slower, show more variation in F0, and less vowel reduction than spontaneous speech. Similar acoustic effects were reported by Dellwo *et al*. (2015) for Zürich German. More importantly, the latter two studies also found that speakers differed in how they adapted their speech between the read and spontaneous styles (Laan, 1997; Dellwo *et al*., 2015, Table 1), thus demonstrating individual differences.

Because a speech sound's linguistic position co-determines its realization, differences in the available speaker-dependent information are expected between different realizations of the same segment, within one speech style. For instance, a consonant in initial, prosodically-strong positions is strengthened in its production relative to that same consonant in non-initial, prosodically-weaker positions (e.g. Fougeron and Keating, 1997). This yields differences in, for example, closure (or linguo-palatal) contact duration during articulation, and such articulatory differences may in turn alter speech sound acoustics. Recently, Smorenburg and Heeren (2020) showed that speaker classification of fricatives /s/ and /x/ was better with tokens sampled from coda rather than onset positions. Moreover, that study demonstrated that the amounts of between- and within-speaker variation depended on syllabic position (see also He and Dellwo, 2017). Building on this earlier work, the current study investigated how the sampling of tokens of the vowel [aː] from different word classes influences the availability of speaker information.

B. Word class

Content words bring richer semantic content to a phrase (i.e. nouns, verbs, adjectives, and adverbs), whereas function words contribute to the phrase's grammatical structure (prepositions, pronouns, auxiliary verbs, etc.). Even though empirical evidence is limited to a handful of studies, these consistently show that whether a token is a content or function word, influences its realization.

Bell *et al.* (2009), amongst others, found that the durations of function words were shorter than those of content words in conversational speech. Moreover, whereas both higher word frequency and word repetition shortened content words, function word duration was not affected by these factors. Studies that investigated the realization of individual segments by word class in read speech found that duration was longer and intensity was higher for the same English vowel /ʊ/, when realized in content relative to function words (Shi *et al.*, 2005), and that a variety of Dutch vowels were more centralized and shorter when pronounced in function words than content words (Van Bergem, 1993, p. 38-39). Because of the systematic variation in vowel realization as a function of word class, the speaker information contained by the same speech sound may be affected by being sampled from a function versus content word.

Function and content words may also differ in phonological properties. For instance, English content but not function words always contain a strong syllable (Selkirk, 1996). For function words, this is only the case when produced in isolation, at the right edge of a major phonological phrase or in focus. A similar pattern is expected in a language like Dutch, which is studied here. Strong syllables carrying word stress are the typical landing sites for pitch accents in Dutch (Sluijter and Van Heuven, 1996), which is why differences in fundamental frequency may be expected between content and function words. These characteristics of function and content words will be considered as confounding factors in this study.

C. Research questions

To further investigate the interaction of linguistic and indexical information, the main research question in the present work is whether word class, i.e. function versus content words, affects the speaker-dependent information carried by the Standard Dutch vowel [aː]. This study thus contributes to understanding if and how sources of variation relevant to voice modelling may vary with linguistic context, and how token sampling may affect acoustic-phonetic FSC. The vowel [aː] was chosen, because it is the most speaker-specific of the corner vowels in Dutch (Van den Heuvel, 1996).

The research question was addressed using data from two corpora, one containing face-to-face conversational speech and one containing telephone conversations. These corpora represented both wide-band (face-to-face) and narrow-band (telephone) recordings, which broadened the evidence base by examining the same effect in two independent speech collections. Moreover, conversational speech, especially when recorded over the telephone, is relevant for forensic application of the results. Note, however, that only non-contemporaneous recordings were available, thus potentially over-estimating the validity of results (Enzinger and Morrison, 2012). A word class effect, however, may be least-confounded in this type of recording because it allows for a direct comparison of tokens from either word class. Moreover, even though background noise was not strictly controlled in these recordings, real forensic data are fully uncontrolled.

To establish that the word class effect on vowel acoustics is present in Dutch spontaneous conversational speech, and not only in lab speech (Van Bergem, 1993; Shi *et al.*, 2005) or the acoustic variable duration (Bell *et al.*, 2009), the word class effect was assessed first in both databases in a control experiment. The main question regarding speaker-specificity was subsequently addressed. The hypothesis was that word class affects the speaker-dependent information contained by the vowel [aː]. This prediction is non-directional, as changes in acoustics related to increased articulatory precision in content relative to function words may help or hinder speaker-dependent information. On the one hand, it has been argued that more precise articulation results in smaller within-speaker variation, which may enhance speaker-specificity (but see McDougall, 2006, fig. 3, for variation in this reduction between speakers). Content words may also facilitate reliable acoustic analysis, because syllables produced with more effort may yield longer segments with a higher signal-to-noise ratio. On the other hand, it has been argued that most speaker-dependent information is found when there is no or a less strict need to attain specific articulatory targets, here: function words. When speakers may adhere more to their own articulatory patterns (see e.g., He and Dellwo, 2017; He *et al.*, 2019), this enlarges between-speaker variation, and as a consequence alters speaker-specificity. As mentioned above, however, speaker-specificity relates between-speaker variation to within-speaker variation. Smorenburg and Heeren (2020) found that the ratio of between- to within-speaker variation was higher for those acoustic-phonetic features that yielded higher speaker classification results. Both types of variation were therefore also measured in the current investigation, as a function of word class. Moreover, in order to study the relationship between acoustic realization by word class and the speaker-dependent information carried by those differential realizations, highly similar acoustic-phonetic features were used in both the control and the main experiment. This choice reduces the maximally obtainable speaker-discriminatory power, but allows for a direct comparison of linguistic effects with indexical information.

Finally, as corpus data were used in the present study, rather than lab or read speech, there are potential confounds to the effect under study. Corpus data were preferred because of their ecological validity, i.e. its representativeness of daily communication and relative closeness to the speech style found in forensic investigations. An effect of word class may be confounded (i) with lexical frequency, i.e. function words tend to be of higher frequency than content words (e.g., Bell *et al.*, 2009), (ii) with phrasal position, i.e. final positions are subject to boundary effects (e.g., Cambier-Langeveld, 2000) and may be more frequent in one word class than the other, and (iii) with pitch accents, as content words, but not function words, are their typical landing sites. In Dutch, pitch accents occur in contents words only if they land in a focused position. These confounding effects were tested as part of the control experiment by labelling [aː] tokens for word frequency, position and the presence/absence of a pitch accent, and assessing the influence of these effects in linear mixed-effects models.

II. METHOD

A. Materials

Spontaneous conversations were taken from the Spoken Dutch Corpus (Oostdijk, 2000). The full corpus consists of fifteen components, covering different speech styles, such as read and conversational speech. Here, two components of spontaneous conversational speech were used, one containing face-to-face speech, and one containing telephone speech recorded over a switchboard. The former sub-corpus contains over 1.7 million words of spontaneous Standard Dutch speech in 925 wave files (stereo recording, 16 kHz sampling frequency), and the latter contains 0.7 million words in 358 wave files (stereo recording, 8 kHz sampling frequency). From each of these two sub-corpora, speech from 50 male, adult speakers of Standard Dutch (aged 18-50) was included. For both types of recordings, speakers were located in their home environments. Interlocutors were instructed to talk for about ten minutes on any topic. For these materials, human-generated orthographic transcripts were available, and using these, additional annotation layers were added to the audio files, containing information on: (a) phonemic content, (b) word class, and (c) word frequency.

To arrive at the phonemic content from the orthography, automatic phonetic transcripts were created through a script using built-in functionality in Praat (Boersma and Weenink, 2018). The resulting phonetic transcript was not error-free, but useful to facilitate the manual selection of vowel tokens (see II.B). Part-Of-Speech (POS) tags were assigned manually to avoid errors, e.g., when one word form has multiple potential POS tags, as in *laat*-AUX ‘let’ vs *laat*-ADJ ‘late’. POS tags were then used for word class labelling into content versus function words. Word frequency information was taken from SUBTLEX-NL (Keuleers *et al.*, 2010), using its POS-specific log10 word frequency. For the face-to-face speech, 9.0% of tokens could not be labelled for frequency, and for the telephone speech, 7.8% were not labelled.

B. Segmentation procedure

Using the automatically-generated phonemic transcripts and speaker metadata, instances of the vowel [aː] produced by the 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 (i) misidentifications of [aː] by the automatic phoneme assignment (e.g., written *a* in English loan words pronounced as [ei] rather than [aː]), (ii) strong reduction or assimilation, where [aː] was not audible or its phonemic nature altered (e.g., *allemaal* 'all' pronounced as /ɑməl/ instead of /ɑləmal/), (iii) background noise or an interfering talker, (iv) hesitations or false starts in the token-bearing word, or (v) interfering sounds by the speaker, such as laughter. If necessary, the automatically determined vowel onset and/or offset locations were adjusted by hand. Using a default range for formant analysis in males (3 formants in 3 kHz), Praat's formant tracks were visually checked against the spectrogram and the analysis range was manually increased or decreased for formant estimation when needed. In total, 3,128 spontaneous face-to-face tokens (1,347 content, 1,780 function words) were manually segmented for 50 speakers (median of 58 tokens per speaker, ranging from 28 to 100+ tokens), and 3,136 spontaneous telephone tokens (1,404 content, 1,732 function words) were manually segmented for another 50 speakers (median of 62 tokens per speaker, ranging from 54 to 100+ tokens).

C. Acoustic analysis

Two types of acoustic variables were extracted from each [aː] token: (i) variables that are expected to vary with word class (and its confounds) based on earlier phonetic research, and (ii) variables that are commonly used in acoustic-phonetic forensic speaker comparisons. Acoustic-phonetic variables were chosen to tie in with the existing linguistic-phonetic literature and to capture their direct effect on speaker-dependent information.

Per [aː] token F0, F1, F2, duration, and intensity were measured. These measures were complemented with formant bandwidth measurements, which may convey articulatory differences between speakers due to their relation with vocal tract tension (e.g., Laver, 1980, ch. 4). Even though the telephone band may affect formant measurements, the F1 of [aː] remains unaffected (Künzel, 2001). All measurements were taken using Praat (Boersma and Weenink, 2018). Segment duration was measured from the manually set onset and offset per token. F1 and F2 were computed (in Hz) using the Burg method (Childers, 1978, pp. 252-255) over the mid 50% of the vowel's duration, as this interval was expected to be minimally influenced by co-articulation. Over the mid-vowel interval, F0 (in Hz) was also measured, using an autocorrelation method. Mean intensity, measured (in dB) as the overall RMS amplitude of the vowel, was determined over the vowel's entire duration, from onset to offset. Intensity was normalized by speaker (z-transforms) to reduce confounding effects of recording conditions.

Polynomial fits of F1 and F2 tracks not only capture resonances at the centre of a vowel, but also transitions in the course of the vowel's duration. These have been shown to carry speaker-dependent information (e.g. Ingram *et al.*, 1996; McDougall, 2004; Morrison, 2009a). The formants were therefore also measured at nine equidistant steps within the vowel (at 10–90% of its duration, window size: 25 ms) and a cubic polynomial fit of these series of measurements was determined per token, using the *poly*() function in R. Per token, this resulted in four coefficients per formant ($f = a_0 + a_1x + a_2x^2 + a_3x^3$), where a_0 captures static formant information in the intercept, and the other coefficients capture the dynamics. The R^2 values for model fit on average were 82% for face-to-face and 81% for telephone speech.

D. Statistical analysis

This section first describes the analysis for the control experiment, which establishes acoustic differences between [aː]s sampled from content versus function words. Next, it presents the analyses run to investigate speaker-dependent information by word class.

1. Linear mixed-effects models

To investigate if word class affects the vowel's acoustic realization linear mixed-effects modelling was used, through the *lmer()* function from the *lme4* package (Bates *et al*., 2015) in R (R Core team, 2016). This was done for each acoustic measure separately (F0, formants, intensity, duration). Significance was evaluated through model comparison using log-likelihood testing; only effects improving the model in a forward-stepwise process were kept in the final model. Models included by-speaker and by-word random intercepts, and the effect of extending the random structure through the addition of by-speaker slopes on final model fit was assessed. A significant contribution from by-speaker slopes would show that speakers differ in how they implement the word classes. Because of the multiple models per data set, a Bonferroni correction was applied to the p-values (.050/5 = .01), and Word Class was binary-coded (content = 0, function = 1). Model fit was checked through examination of the residuals, and this showed that F0 needed to be transformed to 1/F0 and durations by log-10.

Three potential confounds were also tested for all acoustic predictors. First, the effect of including Word Frequency as a factor in the linear mixed-effects models was assessed. Second, boundary effects on [aː] realization were checked by coding if a vowel was realized in the phrase-final word or not. If the effect of Word Class would alter in case a word was produced in non-final position only, this would be indicative of a potential boundary confound in the overall results. Third, the confound of a pitch accent landing on a lexically stressed syllable in a content word was evaluated. Potential pitch accents were acoustically defined as F0 on the target vowel being at least 25 Hz (3–4 semitones) higher than its left and right neighboring syllables. If the effect of Word Class is similar in non-accented and accented vowels, pitch accents resulting from the content word's position in the utterance cannot (fully) explain the results.

2. Measuring speaker-dependency

As measures of within-speaker and between-speaker variation, variances were computed for those acoustic variables showing significant effects in the control experiment. Per acoustic variable, within-speaker variance was computed as the variance by speaker and averaged; between-speaker variance was computed using a leave-one-out approach, thus capturing its variation, and averaged. Through linear mixed-effects modelling (using the same general method as explained in D.1), the effect of Word Class on the two types of variance was assessed.

Next, the effect of Word Class on the available speaker information in [aː] was evaluated in two ways, thus comparing a method from acoustic phonetics to one from forensic phonetics: (i) speaker classification through multinomial logistic regression (MLR), and (ii) the computation of strength-of-evidence using Bayesian likelihood ratios (LRs), respectively.

a. Multinomial logistic regression. MLR is a classifier which estimates regression coefficients per speaker, using as predictors the acoustic variables and the Word Class they were sampled from. To predict speaker identity, the full set of thirteen acoustic predictors was initially included: 1/F0 measured over the mid-50% section of the vowel's duration, the coefficients of the cubic formant fits (the intercepts showed correlations of over r = .97 with the mid-formant measurements), log transforms of the formant bandwidths, log transformed duration, and normalized mean intensity. Correlations between predictors were examined first, and the maximum correlation of $r = -.43$ was not deemed a risk for entering factors together. MLR was implemented in the *multinom* function from the *nnet* package (Venables and Ripley, 2002) in R. The *buildmer* package (Voeten, 2019) was used to automatically determine the optimal model.

The initial, maximal model consisted of all acoustic predictors, the linguistic predictor Word Class, and the first-order interactions of acoustic predictors with Word Class. From this initial model the maximal converging model was determined first, and then the optimal model was fit through backward elimination, using likelihood ratio tests. This was done for both datasets independently: face-to-face and telephone speech.

If the linguistic predictor Word Class was part of an optimal model, likelihood ratio tests were used to compare the model with Word Class to one without it, to thus evaluate its contribution. In case of a significant contribution, speaker-classification accuracy was computed per Word Class by asking the optimal model to predict speaker classifications for tokens from either class. The contributions of the different types of acoustic predictor to speaker classification were assessed by comparing classification performance between the optimal model and the model without a certain predictor type.

b. Likelihood ratio computation. In forensic phonetics, the speaker discriminatory potential of a speech feature can be expressed in terms of the *strength of evidence* (Aitken and Lucy, 2004). This is computed as the likelihood ratio (LR) of two conditional probabilities; the probability of obtaining the evidence while assuming that different speech fragments came from the same speaker, divided by the probability of obtaining the evidence assuming that the different speech fragments came from different speakers. In the case of Forensic Speaker Comparisons, ‘evidence’ is operationalized as the comparison of measurements taken from the two speech fragments. Note that in this study, LRs were used to express the speaker-discriminatory potential of [aː]s sampled from different word classes, not to build a competitive system for use in FSC.

To evaluate the speaker-discriminant potential of [aː], LRs were computed for known same-speaker and known 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. In this investigation, there were 50 same-speaker comparisons, and 1,225 (=[50 × 49]/2) different-speaker comparisons, per database. Because there was only one recording 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, one speaker's first half was compared to a higher-numbered speaker's second half. Relative to speech collections that have multiple recordings per speaker, within-speaker variation may be underestimated here. This should mainly be seen as a restriction on system performance, which may be over-estimated (Enzinger and Morrison, 2012), but not on an effect of Word Class. For the latter, the same recording poses optimal conditions for direct comparison.

LRs were computed, by Word Class and for both speech collections, using three sets of acoustic features. Firstly, only those acoustic variables were included that significantly differed between function and content words in the control experiment: formants (here, their fit coefficients) and duration. Secondly, the same acoustic variables as in the optimal MLR model were used, thus allowing for the most direct comparison with the MLR results. Thirdly, all acoustic variables were included.

To compute LLRs for the multivariate acoustic representation of [aː] tokens, the first step was a sequential leave-one-out (or cross-validated) implementation (see Morrison, 2011) of the method developed in Aitken and Lucy (2004). This method was executed via the MATLAB-script developed by Morrison (2007). The algorithm models within-speaker variance using a normal distribution, and between-speaker variance using multivariate kernel density. Thus, scores for each within-speaker and between-speaker comparison were computed. Next, scores were transformed to LLRs using logistic regression calibration implemented in MATLAB (Morrison, 2009b). For calibration, again a leave-one-out method was used, in which 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. Finally, to avoid extrapolation errors, LRs were limited using an Empirical Lower and Upper Bound (ELUB) LR (Vergeer *et al.*, 2016), computed with one consequential misleading LR[i].

Results of the three feature sets, on either Word Class, were assessed through the median LLRs as well as performance measure C_{llr} (Brümmer and du Preez, 2006).The distance between the median LLR for same-speaker comparisons versus that of different-speaker comparisons is representative of the features' ability to separate the two types of comparisons, and therefore speakers. Along the LLR scale values above 0 represent stronger evidence for the same-speaker hypothesis, whereas values below 0 give stronger evidence for the different-speaker hypothesis. An LLR of 1 means that the evidence is 10 times more likely under the same-speaker hypothesis than under the different-speaker hypothesis, and an LLR of –1 means that the evidence is 10 times more likely under the different-speaker hypothesis. The log-likelihood ratio cost function (C_{llr}) is presented as a performance measure; it not only takes into account the system's correct versus incorrect decisions, but also the values associated with these decisions. It reflects the validity and quality of a system, and the closer to zero, the better.

III. RESULTS

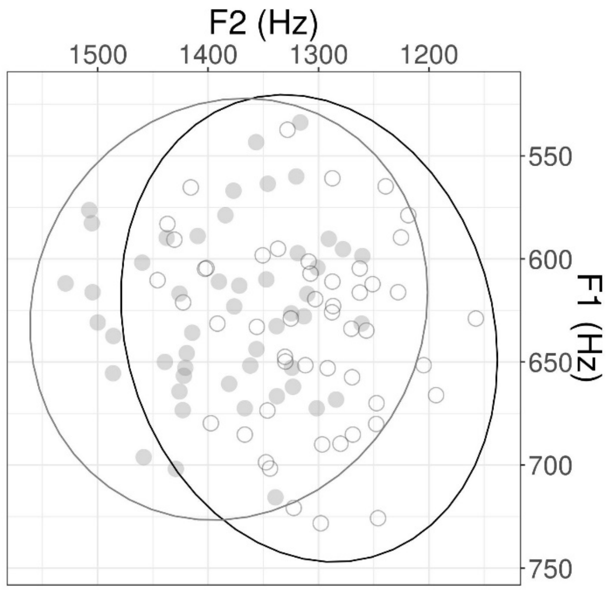


Figure 1: Scatter plot showing F1-F2 means per speaker, for content words (black, open dots) and function words (gray, closed dots). The 95% confidence interval is shown per word class. The plot was created using visiblevowels.org

A. Control experiment: word class effect on acoustics

Figure 1 shows the mean formant frequency values for each speaker plotted in the F1 by F2 plane for content and function words in face-to-face speech. As can be seen, vowel realization partly depends on word class, as confirmed by the statistical analyses. For each acoustic variable in the control experiment, the final mixed-effects model's coefficients are given in Table I. The left half of the table presents results for face-to-face speech (*f2f*, N = 3,128), the right half for telephone speech (*tel*, N = 3,136). The factor Word Class was included in the final models for F1 (*f2f*: $\chi2(1) = 10.6$, $p = .001$; *tel*: $\chi2(1) = 14.5$, $p < .001$), duration (*f2f*: $\chi2(1) = 19.7$, $p < .001$; *tel*: $\chi2(1) = 44.3$, $p < .001$), and it marginally contributed to F2 (*f2f*: $\chi2(1) = 5.5$, $p = .019$; *tel*: $\chi2(1) = 6.1$, $p = .013$). Taken together, results reflected that in function relative to content words the F1 of [a:] was decreased, the F2 was marginally increased, and duration was shorter.

In both speech collections, final models contained by-speaker slopes for Word Class for duration (*f2f*: $\chi2(2) = 9.6$, $p = .004$; *tel*: $\chi2(2) = 27.3$, $p < .001$), and intensity (*f2f*: $\chi2(2) = 14.7$, $p < .001$; *tel*: $\chi2(2) = 25.6$, $p < .001$). In telephone speech, by-speaker slopes also improved the F0, F1, and F2 models (F0: $\chi2(2) = 18.2$, $p < .001$; F1: $\chi2(2) = 62.3$, $p < .001$; F2: $\chi2(2) = 46.4$, $p < .001$).

TABLE I: Linear mixed-effects modelling results for the two data sets, showing significant model coefficients with their corresponding standard errors between parentheses.

	Face-to-face conversation		Telephone conversation	
	β_0	β_1	β_0	β_1
Variable	intercept	word class*	intercept	word class*
F1 [Hz]	640.5 (6.2)	–22.7 (7.0)	677.9 (7.0)	–28.1 (8.0)
F2 [Hz]	1308.1 (10.4)	27.3 (11.6)	1348.9 (11.6)	27.1 (11.9)
F0 [1/Hz]	0.0085 (0.00016)		0.0083 (0.00016)	
duration [log(ms)]	–0.952 (0.009)	–0.068 (0.016)	–0.955 (0.010)	–0.124 (0.020)
intensity [dB]	66.9 (0.7)		67.1 (0.7)	

* reference level = content words

As for the analysis of confounding effects, the addition of the factor Word Frequency did not change model fit, and was therefore not maintained in any of the optimal models. With respect to boundary effects, when only non-final realizations were included in modelling, all differences between content and function words were maintained in both datasets, and in the same direction. As regards a pitch accent confound, all word class differences were maintained when pitch-accented tokens were excluded. In both cases, model coefficients were, of course, not exactly the same (see Supplement*, Tables I and II). These outcomes indicate that these confounds do not affect the word class results as presented in Table I.

Using speech from two independent datasets, a systematic effect of Word Class on [aː] realization was found. In accord with results on Dutch read speech, vowel duration was longer and formant values were less centralized in content than function words (Van Bergem, 1993, p. 34, 39). Intensity and F0 did not vary by word class. Finally, by-speaker slopes in the modelling of several acoustic variables indicated differential pronunciation adaptation to word class between different speakers, especially in the telephone speech collection. With variation in the realization of [aː] by word class established, combined with individual differences in this variation, the next step was to examine speaker-discriminatory information by word class.

B. Speaker-dependency: variances

Using linear mixed-effects models, within-speaker variances were compared between word classes, for those acoustic variables that were significantly different in the control experiment: F1, F2 and duration. The same was done for between-speaker variances.

In both speech collections, within-speaker variances were smaller in content words than function words, for duration and F2 (*f2f*, F2: $\chi2(1) = 12.7$, $p < .001$, duration: $\chi2(1) = 17.4$, p

<.001; *tel*, F2: χ2(1) = 4.7, p = .03, duration: χ2(1) = 34.3, p <.001). The variances can be found in the Supplement* (Table III), but as an example: when looking at the within-speaker variability in the F2 model for face-to-face speech, content words had a 67.3 Hz smaller standard deviation[ii] than function words.

In both speech collections the between-speaker variance was larger for all variables in function than content words (*f2f*, F1: χ2(1) = 237.3, p < .001, F2: χ2(1) = 527.5, p < .001, duration: χ2(1) = 680.2, p <.001; *tel*, F1: χ2(1) = 363.5, p < .001, F2: χ2(1) = 372.4, p < .001, duration: χ2(1) = 676.2, p <.001). For example, the between-speaker standard deviation in the F2 model for face-to-face speech was 71.2 Hz smaller in content words than function words.

For all other acoustic-phonetic measures both within- and between-speaker variances showed the same trend of reduced size in content words (see Supplement*, Table III).

C. Speaker-dependency: MLR results

For face-to-face conversation (*N* = 3,128), the optimal MLR speaker-classification model included the predictor Word Class (χ2(637) = 1216, p < .001); classification performance was 32.1% correct on content words, and 29.3% on function words (chance level ≈ 2%). The model also contained formant (bandwidth) information (except fit coefficient a_3 for F1), F0, duration, and intensity, and all acoustic predictors also interacted with Word Class. The order in which predictors contributed most to classification performance was: formants, F0, intensity, and duration, with respective reductions in maximal classification performance from 30.7% to 10.7%, 24.3%, 28.1% and 29.0%, when the predictor was left out. Leaving out either formant intercepts (a_0) or dynamic formant information (a_1, a_2, a_3) gave performance reductions from 30.7% to 22.6% and 26.4%, respectively.

Also for telephone speech (N = 3,136), the optimal speaker model included Word Class ($\chi 2(490) = 1186$, $p < .001$); speaker classification for content words was 24.0% correct, whereas for function words it was 21.5% correct. The model furthermore contained the formant coefficients (except fit coefficient a_3 for F2), F0, and duration, and these acoustic predictors also interacted with Word Class. Not included were formant bandwidths and intensity. The order in which acoustic predictors contributed most to speaker classification was: formant coefficients, duration and F0, with respective reductions in maximal classification performance from 22.6% to 8.8%, 15.6% and 17.7%. Leaving out either formant intercepts or the higher coefficients yielded performance reductions to 14.2% and 18.0%, respectively.

D. Speaker-dependency: LR results

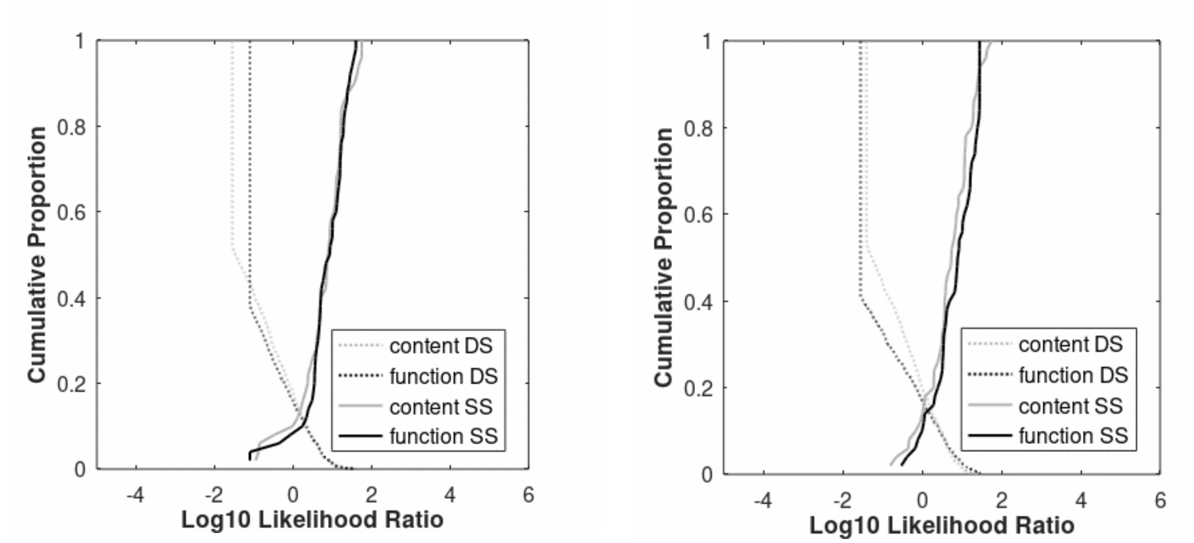
The median log-likelihood ratios and C_{llr}s for [aː]s sampled from either Word Class are given in Table II, for each of the three acoustic feature sets separately (see section II.D.2.b). Median LLRs were computed for same-speaker comparisons (LLR_{SS}) and for different-speaker comparisons (LLR_{DS}).

When comparing between the word classes, per feature set, median LLRs are close together. LLR_{SS} tend to be slightly more positive for function than content words in both speech collections, whereas LLR_{DS} show this trend for some feature sets, but the opposite trend in others. However, the order of magnitude of the LLRs remains comparable between word classes. For face-to-face speech, LRs do not improve when the MLR feature set is extended to all acoustic-phonetic variables, whereas they do in telephone speech. Remember, however, that for telephone speech, the MLR predictor set was smaller than for face-to-face speech. This makes the difference between the MLR- and all-feature sets larger in telephone speech. The general trend in Table II is that performance improves with the number of acoustic features included.

TABLE II: Results for face-to-face (f2f) and telephone (tel) speech, for either content (N_{f2f} = 1,443; N_{tel} = 1,318) or function words (N_{f2f} = 1,492; N_{tel} = 1,617), showing median LLR for both same-speaker and different-speaker comparisons, and C_{llr}.

data	feature set	word class	Md(LLR_{SS})	Md(LLR_{DS})	C_{llr}
f2f	formants, duration	content	0.41	–0.28	0.850
		function	0.42	–0.34	0.814
	as in MLR	content	0.90	–1.47	0.590
		function	0.94	–1.10	0.600
	all	content	0.91	–1.43	0.594
		function	0.99	–1.10	0.597
tel	formants, duration	content	0.68	–1.05	0.665
		function	0.70	–1.55	0.593
	as in MLR	content	0.74	–1.27	0.636
		function	0.92	–1.55	0.561
	all	content	0.96	–1.25	0.550
		function	1.10	–1.25	0.526

When looking at the C_{llr}s the pattern of results seems somewhat different for face-to-face than telephone speech. In the latter speech type, function word [aː]s do somewhat better than content word [aː]s. In face-to-face speech, the relation between function and content word C_{llr}s varies by feature set. To illustrate the comparable behavior between the word classes Figure 2 shows Tippett plots for both collections using results from the MLR feature set.



a. b.

Figure 2: Tippett plots for LR results based on the MLR feature sets, showing both same-speaker (SS, solid line) and different-speaker (DS, dashed line) LLRs. In (a) face-to-face speech and (b) telephone speech performance is compared between content words (gray) and function words (black).

With comparable results for the two word classes, a post-hoc analysis was done using data mixed between word classes, thus allowing for more data to be included in the computation of strength-of-evidence. LRs were computed including both word classes per speech collection and using all acoustic variables, i.e. the best-performing feature set. For face-to-face speech, the median LLR_{SS} was 1.0 and the median LLR_{DS} was –1.16, with the C_{llr} at 0.616, which gives similar discriminatory power and a slight reduction in performance relative to one word class only. For telephone speech, the median LLR_{SS} was 1.33 and the median LLR_{DS} was –1.7, with the C_{llr} at 0.429. Here, the mixed condition shows some improvement over the individual word classes.

IV. GENERAL DISCUSSION

This study investigated if speaker-specific information carried by the Standard Dutch vowel [aː] varies with the word class tokens are sampled from. Using conversational speech from two corpora, face-to-face and telephone speech, it was first established that vowel realization in conversational speech varies by word class along multiple acoustic dimensions, as in lab speech (Van Bergem, 1993; Shi *et al.*, 2005). As expected, spectral and temporal vowel reduction in function words resulted in more centralized positions of the vowels in the acoustic space and shorter durations than in content words. Such differential acoustics would potentially yield differences in the speaker information available per word class. Therefore, the main experiment addressed the question of whether the word class from which [aː] samples are taken affects their amount of speaker-dependent information conveyed.

Results showed that word class impacted both within- and between-speaker variation, but that the effect of word class on speaker separation was not fully consistent across the two speaker modelling approaches. The vowel [aː] yielded somewhat better speaker-classification scores in content than function words, in both speech collections, whereas the strength-of-evidence derived from the same acoustic feature set did not reflect this difference. What both analyses agreed on, however, was that there is speaker-dependent information in just the vowel [aː] when sampled from spontaneous (telephone) speech. This adds to earlier acoustic-phonetic work on speaker-dependent information in vowels conducted on less spontaneous materials, e.g. read speech (e.g., McDougall, 2006; Morrison, 2009a), or semi-spontaneous speech (Gold, 2014, ch. 5; Rose, 2015).

Speaker classification through MLR showed a small, yet consistent, benefit of content over function words on the speaker information contained by [aː], whereas LRs showed results that were comparable for both word classes. This discrepancy between the methods must be explained by differences in the modelling between them. LRs take into account both within-speaker and between-speaker variation. It is not surprising that LRs are comparable for the two word classes, when considering that the ratio of between-to-within speaker variances remained comparable between content and function words; when one type of variance increased, the other one did as well, and vice versa. MLR results are well-explained when taking into account either within-speaker or between-speaker variation. Comparable statistical techniques have yielded results consistent with the word class effect obtained here (McDougall, 2004; He and Dellwo 2017; Smorenburg and Heeren, 2020). On the one hand, the more precise articulation in content as opposed to function words, as reflected by smaller within-speaker variation, is in line with a speaker-classification advantage in read speech for nuclear-stressed versus non-nuclear-stressed syllables (McDougall, 2004). In that study, Linear Discriminant Analysis (LDA) was used for speaker classification. At the same time, more between-speaker variation was here found in function than content words, that is in contexts with less strict articulatory demands. This has been reported before by e.g., He and Dellwo (2017), who investigated between-speaker variation in intensity contours in the opening versus closing gestures of a syllable. Using MLR modelling, they found that measures taken from that part of the syllable which presumably has less strict articulatory targets, i.e. the second half of a syllable, accounted for most between-speaker variation. Recently, similar effects were demonstrated for F1 dynamics, which contained more between-speaker variation in closing than opening gestures (He *et al.*, 2019), and for Dutch fricatives /s, x/ showing more between-speaker variation in codas than onsets (Smorenburg and Heeren, 2020). The results from the current investigation suggest that speaker classification models, such as MLR and LDA, do not use within- and between-speaker variation in the same way for speaker modelling as the forensic standard, LRs, does.

Recall that speaker-specific features for FSC ideally exhibit small within-speaker variation combined with large between-speaker variation. As the two types of variance were found to co-vary in size with word class, differences in speaker-specificity by linguistic condition were minimized in LR computations. Therefore, while acoustic-phonetic research into individual differences and context-dependent variation within and between speakers is crucial for understanding speech communication, the *speaker-specificity* of speech features may be best-captured by the reporting standard of the court, i.e. the LR approach. The relevance of both within- and between-speaker variation for speaker separation is furthermore consistent with voice perception models (Lavan *et al.*, 2018). What the current results add to the existing literature is the consideration that the amount of variation displayed within and between speakers may depend on the linguistic context from which samples are taken. Models of voice perception take a prototype-based approach, where it is assumed that unfamiliar voices are processed as deviations from the prototype, whereas familiar voices are recognized as patterns without reference to the prototype (see Kreiman and Sidtis, 2011, ch. 5). Especially for the recognition of unfamiliar speakers, linguistic conditions affecting the size of variances may affect the deviation from the prototype and thus yield differential performance.

In both MLR and LR modelling various acoustic predictors contributed speaker information. The predictors that carried most information were spectral in nature: formants’ averages, dynamics, and –to some extent– their bandwidths. This was most evident from the speaker classification results, but is also reflected by comparing LR results between feature sets. This finding ties in with earlier research on speaker-dependent information in vowel formants (e.g. McDougall, 2004, 2006), and is in line with the finding by Bachorowski and Owren (1999) that within a group of same-sex speakers, as used in the current investigation, vocal-tract variables are more informative than the vocal source variable. In the MLR model for face-to-face speech, formant bandwidths were also kept, suggesting that they carried speaker-dependent information, which – to the author's knowledge – is a first demonstration; their contribution may be explained by the fact that bandwidths reflect between-speaker differences in vocal tract tension (Laver, 1980, ch. 4). Duration and intensity held little speaker information. Duration is strongly influenced by speech tempo (Van den Heuvel, 1996, p. 77), and this – when measured as articulation rate – contains relatively little information as a speaker discriminant (Quené, 2008; Gold, 2014). Intensity is likely to be influenced by the recording conditions, especially when spontaneous speech is collected under naturalistic conditions as the data used here, and probably even more so when uncontrolled recordings are involved as in forensic casework.

Focusing on the formants, earlier studies have reported that dynamic representations of formant trajectories carry speaker-dependent information (e.g., Ingram *et al.*, 1996; McDougall, 2006; Hughes *et al.*, 2016). In the present study, this was also reflected by the MLR results, but dynamic formant information, as captured by the higher fit coefficients, contributed less than static formant intercepts. One reason why the contribution of formant dynamics may be restricted is that the Dutch vowel /aː/ is not a diphthong, thus containing little inherent transition that may yield articulatory differences between speakers. In several earlier studies, diphthongs or segmental combinations were used (e.g., McDougall, 2004; 2006; Morrison, 2009a). In a study on the speaker-dependency of hesitation markers sampled from British English spontaneous speech (i.e. with varying contexts), formant dynamics only aided in *um*, with inherent vowel-to-consonant transition, not in *uh*, without transition (Hughes *et al.*, 2016). However, Rose (2015) found stronger speaker evidence with formant trajectories than mid-vowel measurements only for steady-state vowel /ɜ/, using samples from eight different word contexts in map task recordings. Another reason for the absence of a more prominent formant dynamics result may be that the variable phonetic contexts in the present investigation reduced their information value, i.e. dynamics were partially determined by neighboring sounds that differed between tokens.

The current results, based on acoustic-phonetic features in vowels in spontaneous speech, tend to show lower LRs than similar studies in the literature (Gold, 2014: table 5.4; Hughes *et al.*, 2016). This difference may be partially explained by the larger effects of co-articulation and contextual variation for [aː] tokens sampled from a large variety of words than for schwa sampled from hesitation markers only (Hughes *et al.*, 2016). In addition, the use of ELUBs in the current study strongly limited the range of accepted LRs, whereas earlier work often did not apply these limits. In comparison with ASR approaches to vowel data, LRs are much lower here; ASR systems use speech features that generally have a higher discriminatory power, such as MFCCs or ivectors. However, in order to investigate the effect of word class acoustics on a vowel's speaker-specific information in a way that ties in with earlier linguistic-phonetic work, the current experiments were intentionally restricted to one vowel and its acoustic-phonetic variables. In FSC casework, acoustic-phonetic analysis includes different aspects of speech (e.g., various segments, intonation, tempo), thus potentially yielding a higher discriminatory power due to their complementarity. If case data and legislation allow, ASR might be used as an additional or even alternative method. What the current results contribute, however, is that the sampling of vowel tokens for acoustic-phonetic FSC, and perhaps also for ASR, is unlikely to depend on the word class from which tokens are sampled.

In this study, LR results (both median LLRs and C_{llr}s) were somewhat better on narrow-band telephone than broadband face-to-face speech. This is considered unexpected, but there are multiple factors that may have contributed to this result. First, the set of speakers differed between speech collections, meaning that the composition of the 50 speakers per database may have affected the outcome. Speakers are known to differ in discriminability by humans (e.g., Baumann and Belin, 2010) and by machines (Doddington *et al.*, 1998), so there may be a sampling effect. Evidence for this is found in the larger number of random slopes in the telephone speech models, which reflects higher between-speaker variation (see III.A). Second, speaking behavior varies by speech style (Moos, 2010; Dellwo *et al.*, 2015), and specifically behavior during telephone conversation may be hypothesized to differ from that in face-to-face speech as speakers are unable to see each other. It is thinkable that speakers therefore articulate relatively clearly in comparison with face-to-face speech, which may aid their discriminability. This explanation is supported by a tendency for smaller within-speaker variances in the telephone speech relative to the face-to-face speech collection (see Supplement). For MLR models, optimal performance on face-to-face speech was better than on telephone speech, but recall that the optimal models for the two collections differed in predictor sets: the former speech type had a larger set of predictors.

For acoustic-phonetic forensic voice comparisons it is important to not only know which features convey most speaker information, but also if it matters where the features are sampled from. The current study shows that even though there are effects of word class on vowel realization and on within- and between-speaker variances of acoustic-phonetic variables, these differences do not affect the strength of evidence contained by [aː]. In casework, there thus seems no principled reason to carefully balance sampling from different word classes or to use one class only, when vowel quality is decisive in the inclusion of tokens (whereas generally, more reduced tokens are expected in function that content words). It remains advisable, however, to be aware of strongly unbalanced sampling across word classes, as they influence the measurement outcome of variables bearing speaker information. Moreover, the present study included speech data with some characteristics also found in casework, but certainly not all. For instance, the collections used here did not contain non-contemporaneous data, and the demographic background of the speakers was not specifically selected. Only age, sex and the use of Standard Dutch were controlled for. This is a limitation, as it is expected to yield a degree of mismatch with speakers encountered in actual casework, however various they may be. Moreover, though male speakers are more prevalent in forensic-phonetic casework, female voices are encountered as well, but they were not part of this study. Although the values of their acoustic measurements are expected to differ from those of males (duration: Quené, 2008; Bell *et al.*, 2009; formants: Adank *et al.*, 2004), no fundamental differences in the interaction between word class and speaker-dependent information are expected between male and female speakers. Finally, this study was restricted to the most speaker-specific vowel in Dutch, [aː]. As differences in vowel realization by word class are not expected to be larger for other vowels of Dutch (van Bergem, 1993), the effect is predicted to transfer to the other vowels. Other linguistic contexts, however, may affect other acoustic variables and thus impact speaker-dependent information differently.

V. CONCLUSION

Not only speech sound or speech style matters as to how much speaker information is available, but – to some degree – also the class of word in which a speech sound is located. Using two independent databases of conversational speech, analyses showed that [aː] acoustics vary with the word class the vowel is realized in, and that [aː] contains less within-speaker variation in content than function words, but also less between-speaker variation in content than function words. Even though this results in slightly better speaker classification for content words, the forensic strength-of-evidence computed from [aː] was comparable between word classes, presumably because it depends on both types of variation.

ACKNOWLEDGEMENTS

This work was supported by the Netherlands Organization for Scientific Research (NWO VIDI grant 276-75-010). I would like to thank Jos Pacilly for help in scripting, David van der Vloed for discussion on the LR analyses, and Laura Smorenburg, Meike de Boer, Cesko Voeten and anonymous reviewers for constructive feedback on earlier versions of this manuscript.

* See supplementary material at [***URL will be inserted by AIP***] for mixed-effect modelling results of the confound analyses, and for within-speaker and between-speaker variances of all variables in the speaker-dependency analysis.

---

[i] ELUBs were computed using an R script developed by the first author of Vergeer *et al.* (2016).
[ii] Standard deviation is given instead of the variance, as the former has an interpretable measurement unit (here: Hertz).

REFERENCES

Adank, P., Van Hout, R., and Smits, R. (**2004**). "An acoustic description of the vowels of Northern and Southern Standard Dutch," J. Acoust. Soc. Am. **116**, 1729–1738.

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

Andics, A. (**2013**). *Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning*. PhD dissertation, Radboud University Nijmegen. Available from https://repository.ubn.ru.nl/handle/2066/101022.

Bachorowski, J.-A., and Owren, M. J. (**1999**). "Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech," J. Acoust. Soc. Am. **106,** 1054–1063.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (**2015**). "Fitting Linear Mixed-Effects. Models Using lme4," J. Stat. Softw. **67**, 1–48.

Baumann, O., and Belin, P. (**2010**). "Perceptual scaling of voice identity: common dimensions for different vowels and speakers," Psych. Res. **74**, 110–120.

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,” J. Mem. Lang. **60**, 92–111.

Boersma, P., and Weenink, D. (**2018**). “Praat: doing phonetics by computer (Version 6.0.42) [Computer program],” http://www.praat.org/ (Last viewed on 1 September 2018).

Brümmer, N., and du Preez, J. (**2006**). “Application-independent evaluation of speaker detection,” Comput Speech Lang **20**, 230–275.

Cambier-Langeveld, G. M. (**2000**). *Temporal marking of accents and boundaries*. PhD dissertation, University of Amsterdam. Available from https://dare.uva.nl/.

Childers, D. G. (**1978**). *Modern Spectrum Analysis*. New York: IEEE press.

Dellwo, V., Leemann, A., and Kolly, M.-J. (**2015**). “The recognition of read and spontaneous speech in local vernacular: The case of Zurich German,” J. Phonetics **48**, 13–28.

Doddington, G., Liggett, W., Martin, A., Przybocki, M., and Reynolds, D. (**1998**). Sheep, goats, lambs and wolves: a statistical analysis of speaker performance. *Proceedings of IC-SLD’98, NIST 1998 Speaker Recognition Evaluation*, Sydney, Australia, pp. 1351–1354.

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

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

Gold, E. and French, P. (**2019**). “International practices in forensic speaker comparisons: second survey,” The Int. J. of Speech, Lang. and the Law **26**, 1–20.

Fougeron, C., and Keating, P. A. (**1997**). “Articulatory strengthening at edges of prosodic domains,” J. Acoust. Soc. Am. **101**, 3728–3740.

He, L., and Dellwo, V. (**2017**). “Between-speaker variability in temporal organizations of intensity contours,” J. Acoust. Soc. Am. **141**, EL488–EL494.

He, L., Zhang, Y., and Dellwo, V. (**2019**). “Between-speaker variability and temporal organization of the first formant,” J. Acoust. Soc. Am. **145**, EL209–EL214.

Hughes, V., Foulkes, P., and Wood, S. (**2016**). Formant dynamics and durations of um improve the performance of automatic speaker recognition systems. In: *Proceedings of the 16th Australasian Conference on Speech Science and Technology* (ASSTA) , University of Western Sydney, Australia.

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

Kavanagh, C. M. (**2014**). *New consonantal acoustic parameters for forensic speaker comparison*. PhD dissertation, University of York. Available from https://core.ac.uk/download/pdf/14343593.pdf.

Keuleers, E., Brysbaert, M., and New, B. (**2010**). “SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles,” Behav. Res. Methods **42**, 643–650.

Kreiman, J., and Sidtis, D. (**2011**). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception* (Wiley- Blackwell).

Künzel, H. J. (**2001**). “Beware of the ‘telephone effect’: the influence of telephone transmission on the measurement of formant frequencies,” Forensic Linguist. **8**, 80–99.

Laan, G. P. M. (**1997**). "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style," Speech Commun. **22**, 43–65.

Lavan, N., Burston, L. F., and Garrido, L. (2018). "How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices," Br. J. Psychol. **110**, S76–S93.

Laver, J. (**1980**). *The phonetic description of voice quality*. Cambridge University Press, Cambridge.

Lee, Y., Keating, P., and Kreiman, J. (**2019**). "Acoustic voice variation within and between speakers," J. Acoust. Soc. Am. **146**, 1568–1579.

McDougall, K. (**2004**). "Speaker-specific formant dynamics: An experiment on Australian English," Int. J. of Speech, Lang. and the Law **11**, 103–130.

McDougall, K. (**2006**). "Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies," Int. J. of Speech, Lang. and the Law **13**, 89–126.

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. (**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. (**2009a**). "Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs," J. Acoust. Soc. Am. **125**, 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, S. G. (**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 Comm. **53**, 242–256.

Morrison, G. S., Sahito, F. H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., and Dorny, C. (**2016**). “INTERPOL survey of the use of speaker identification by law enforcement agencies,” Forensic Sci. Int. **263**, 92–100.

Oostdijk, N. (**2000**). “The Spoken Dutch Corpus. Overview and first evaluation,” *Proceedings of LREC 2000*, Athens, Greece, pp. 887–894.

Palmeri, T. J., Goldinger, S. D., and Pisoni, D. B. (**1993**). Episodic encoding of voice attributes and recognition memory for spoken words. *J. Exp. Psychol. Learn.* **19**(2), 309–328.

Peterson, G. E., and Barney, H. L. (**1952**). “Control methods used in a study of the vowels,” J. Acoust. Soc. Am. **24**, 175–184.

Quené, H. (**2008**). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. J. Acoust. Soc. Am. **123**, 1104–1113.

R Core Team (**2016**). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ (Last viewed on 15 October 2017).

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 International Conference on Acoustics, Speech and Signal processing (ICASSP)*, Brisbane, pp. 4819–4823.

Selkirk, E. (**1996**). The Prosodic Structure of Function Words. In *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, edited by J. L. Morgan and K. Demuth (Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.), pp. 187–213.

Shi, R., Gick, B., Kanwischer, D., and Wilson, I. (**2005**). "Frequency and Category Factors in the Reduction and Assimilation of function Words: EPG and Acoustic Measures," J. Psycholinguist. Res. **34**, 341–364.

Sluijter, A. M., and Van Heuven, V. J. (**1996**). "Spectral balance as an acoustic correlate of linguistic stress," J. Acoust. Soc. Am. **100**, 2471–2485.

Smorenburg, L., and Heeren, W. (**2020**). "The distribution of speaker information in Dutch fricatives /s/ and /x/ from telephone dialogues," J. Acoust. Soc. Am **147**, 949–960.

Van Bergem, D. (**1993**). *Acoustic and lexical vowel reduction*. PhD dissertation. University of Amsterdam.

Van Berkum, J. J. A., Van den Brink, D., Tesink, C. M. J. Y., Kos, M., and Hagoort, P. (**2008**). "The neural integration of speaker and message," J. Cognitive Neurosci. **20**, 580–591.

Van den Heuvel, H. (**1996**). *Speaker variability in acoustic properties of Dutch phoneme realisation*s. PhD dissertation, Radboud University Nijmegen, available from lands.let.ru.nl/literature/theses/heuvel_thesis.ps.

Venables, W. N., and, Ripley, B. D. (**2002**). *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

Vergeer, P., van Es, A., de Jongh, A., Alberink, I., and Stoel, R. (**2016**)."Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating?," Sci. Justice **56**, 482–491.

Voeten, C. (**2019**). buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression. R, package version 0.1. Available from https://cran.r-project.org/ (Last viewed on 18 April 2019).

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 Comm.* **55**, 796–813.