Do Sound Segments Contribute to Sounding Charismatic? Evidence from a Case Study of Steve Jobs' and Mark Zuckerberg's Vowel Spaces

Oliver Niebuhr

University of Southern Denmark, Mads Clausen Institute, Alsion 2, DK-6400 Sønderborg, Denmark.

Simon Gonzalez

Centre of Excellence for the Dynamics of Language, The Australian National University, 13 Ellery Crescent, Canberra ACT 2601.'

(Received 24 June 2018; accepted 22 November 2018)

The paper presents a case study of two popular US American CEOs. It compares the acoustic vowel space sizes of the more charismatic speaker Steve Jobs and those of the less charismatic speaker Mark Zuckerberg, as part of an initial acoustic step to examine a traditional claim of rhetoric that clearer speech makes a speaker sound more charismatic. Analysing about 2,000 long and short vowel tokens from representative keynote speech excerpts of the two speakers shows that Jobs' vowel space is, across various segmental and prosodic context factors, significantly larger than that of Zuckerberg, whose vowel space is strongly reduced particularly when addressing investors. The differences in vowel-space size are consistent with the claim of rhetoric that a clear articulation is a key characteristic of a charismatic speaker. The discussion of the results describes further experimental steps required to back up the link between clear pronunciation and speaker charisma.

1. INTRODUCTION

1.1. The Phonetics of Charismatic Speech

Spoken language is not just the exchange of propositions. On the contrary, it is in the first place a social action, and "this fact both shapes the nature of the activity and its consequences."[26] We use speech for expressing our emotions and sharing them with others, as well as for influencing the thoughts and actions of others. To this extent, charisma as "the art of persuasion"[65] through "emotion-laden leader signalling"[1] is indeed a core element of spoken language — and its phonetic essence is surprisingly little understood.

It is against this background that Rosenberg and Hirschberg called for an empirical definition of charisma in speech.[53] They analysed the acoustic-prosodic characteristics of male US politicians and related them to perceived charisma. Their analysis led to the conclusion that higher levels of fundamental frequency (F0), intensity, and speaking rate, as well as a larger F0 range, make speakers sound more charismatic. These findings were consistent with analyses of other political leaders in the US and in Europe.[5,6,25,61,65] Moreover, the same prosodic strategy also works for business leaders,[42,43] except that females are more likely to lower rather than raise their F0 level.[46] Furthermore, Niebuhr et al. added to the picture that shorter prosodic phrases, larger numbers of emphatic pitch accents, high-energy voices (higher values of %V, spectral emphasis, and HNR[1]), and more variable speech rhythm (higher VarcoV values) also support a speaker's charismatic impact.[42,43,45]

While an empirical definition of charismatic prosody is within reach (at least for Western Germanic languages and/or the Western culture), a whole area of speech has hardly been addressed so far: sound segments.[5] Manuals on rhetoric and leadership have claimed ever since that clear and crisp articulation of "every phrase and word"[40] "is imperative to develop charisma."[7]

Basically, this claim makes sense from the perspective of basic ethological principles like the Effort Code.[21] According to the Effort Code, a fundamental behavioural pattern of all biological organisms, they spend more time and effort on things and actions that are more important to them. In order to understand the implications of this basic principle in human everyday life, one only needs to think about how elaborate the table is set when important guests are coming as compared to how simple the table setting is when one eats alone. If a simple table setting were used for important guests, the implicit message would be that the host does not care about his/her actions and/or that the invited guests are not important to the host. The same is true for speech communication. Investing more effort into articulating sound segments would indicate from the Effort Code's point of view that the conveyed message is important and that the speaker shows appreciation for his/her audience. In contrast, mumbling would implicitly signal that the speaker does not care about his/her message and the audience as well.

Similarly, the Hypo-Hyper (H&H) theory of Lindblom regarded a clear, effortful articulation (hyper-speech) as being listener-oriented, with, for example, an aim to meet the re-

[1] %V is the average proportion of vowel segments in an utterance. It was introduced as a rhythm measure, but since vowels are the most energetic parts of an utterance, %V is also highly correlated with the perceived "volume" and "power" of a voice.[50] Spectral emphasis refers to the difference between the total acoustic energy of the signal at a given point in time and the energy in the lower frequency region of that signal (0-1.5*f0median, following Traunmüller and Eriksson).[66] Thus, spectral emphasis quantifies the loss of energy towards higher frequencies. Voices with a perceived high "volume" and "power" lose less energy towards higher frequencies. HNR stands for harmonics-to-noise ratio and quantifies the ratio (in dB) between the periodic energy and the noise energy of a signal at a given point in time.

quirements of a formal situation, making it easier for the audience to understand the speaker's message.[35] Thus, less clear and effortful articulation (hypo-speech) occurs when the speaker places his/her own interests (e.g. being efficient and saving articulatory energy) above those of the audience (e.g. being intelligible and easy to follow). Obviously, this is the opposite of what charismatic speakers are supposed to do.

The predictions of the Effort Code and the H&H theory manifested themselves in experimental research on how speech becomes more effortful in adverse conditions or for larger speaker-listener distances,[12] how speakers produce and perceive the expression of surprise,[8] and how post-lexical reduction processes among consonants influence a speaker's perceived personality traits. A recent study by Niebuhr found that speakers whose utterances were characterized by a larger number of reduced (e.g., elided, lenited, or assimilated) consonant articulations sound less sociable, educated, and sincere in the ears of listeners.[44] These personality traits were similar to attributes like "upright", "team-integrating", and "charming", which, in turn, were shown to be correlated with speaker charisma.[53,69] Nevertheless, although such results basically argued in favour of the claim that the" clarity of speech and pronunciation is important for perceived charisma and influential delivery",[38] and studies like that of Niebuhr represented no clear, direct evidence for a positive correlation between the clarity of speech pronunciation and the perception of speaker charisma.[44]

Addressing this gap, the present paper represents a second step within a line of research that examines the assumed link between sound segments and their pronunciation on the one hand and perceived speaker charisma on the other. Clarity of pronunciation is obviously related to a higher speech-production effort. But, as the latter is hardly objectively measurable (an issue already brought up in connection with the Hypo-Hyper theory of Lindblom and not solved since then), clarity of pronunciation is defined in our line of research as the level of distinctivity in the (acoustic) phonetic implementation of phonological contrasts and, as one countable feature of this distinctivity, the frequency of speech-reduction processes such as those manipulated in the study of Niebuhr above.[35,44]

A first step in this line of research was made in the studies of Niebuhr et al., based on a comparative case study of two popular US American CEOs: Steve Jobs (SJ) and Mark Zuckerberg (MZ).[45] The following section explains why SJ and MZ represent an ideal pair of speakers to start with.

1.2. Motivation of the Case Study

SJ is well known and often cited for his charismatic speeches,[59] whereas MZ's public-speaking skills made some researchers and journalists, e.g., Tobak,[64] question the relevance of charisma in modern leadership. SJ and MZ are often named as examples of different types of charismatic business leaders, as in the CNN article by Sutter" when it comes to presentation, Mark Zuckerberg is no Steve Jobs."[63] Gruener also compares the two CEOs' public-speaking performances and concludes that "Jobs [...] owns the kind of charisma very few people have — the kind that makes you drop everything you do and listen instantly. Zuckerberg does not have that charisma, not yet, and the presentation skills are rough enough to impact Facebook's perception in a negative way."[20] Similar impressionistic statements were made from many other people whose professional backgrounds range from bloggers to journalists to experts in rhetoric.

Recently, a formal perception experiment further elucidated these public assessments of the two speakers (see Niebuhr et al.).[45] Ninety-eight English-speaking listeners were asked to rate the leadership experience, charisma, and charisma-related personality traits of SJ and MZ based on 30-second keynote excerpts. These excerpts were low-pass filtered (0–500 Hz) in order to remove verbal content and speaker identity from the stimuli. Low-pass filtered speech is also called "delexicalized" speech.[55] What remains in a low-pass filtered signal is F0 and, for most of the time, 1–2 subsequent harmonics, and, on this basis, the first formant frequency (F1).[2] This is sufficient for listeners to still perceive all the variable prosodic patterns of intonation, voice quality, loudness, tempo (based on the syllable-related waxing and waning acoustic-energy alternations), phrasing, and durations of utterances and pauses. What a low-pass filtered signal lacks is the critical region between about 500–3,000 Hz of the speech signal (the typical telephone transmission bandwidth, cf. Miet et al.) in which most information about the types of sound segments and their places of articulation are encoded in terms of F1, F2, and F3.[39] It also lacks the amplitudes and frequencies of formants higher than F3 that "are greatly related to the anatomy of one's laryngeal cavity [...] and therefore carry some individual specificity."[71] F0 and the spectral frequency information up to 500 Hz are only related to speaker identity in that they can to some degree indicate group features like speaker sex, age, and (non)native language background. However, individual speakers cannot be identified on this basis.

The 30 second stimuli of SJ and MZ were embedded in a larger set of equally long keynote excerpts of other male English keynote speakers and presented to the 98 listeners in individually randomized orders. Results showed that the 98 listeners gave SJ significantly and substantially higher ratings than MZ on all scales. Figure 1 summarizes the relevant results of the perception experiment.[3]

The results in Fig. 1 put the informal observations of individual journalists, bloggers, and rhetoric experts on a broader, systematic, and therefore more objective empirical footing. That is, that SJ was deemed as more charismatic than MZ did not just reflect the opinion of individual listeners. Rather, the opinion was of a general nature. Additionally, since speaker iden-

[2]Formant is a technical term in the phonetic sciences for the resonance frequencies that the pharyngeal, oral, and nasal tubes and cavities impose on all source signals generated in the vocal tract for speech communication. The vowels and consonants of the world's languages are particularly shaped by the two lowest formant frequencies, F1 and F2. In vowels, F1 can vary from 300 Hz to 1000 Hz. It is correlated with the vertical tongue position. The lower F1 is, the higher is the vowel, i.e. the more the tongue is raised towards the roof of the mouth (/i/ has a low and /ɑ/ a high F1). F2 can vary from 550 Hz to 2500 Hz. F2 values indicate the horizontal tongue position, i.e. the frontness or backness of a vowel. A vowel with high a F2 value is articulated in the front of the mouth. A back vowel has a low F2 value (/i/ has a high and /ɑ/ a low F2). Lip rounding lowers F2 and F3 as well. Higher formants such as F4 and F5 are more strongly associated with voice qualities than with sound qualities.[28]

[3]However, also note that MZ is by no means an uncharismatic speaker. He is merely *less* charismatic than SJ in the analysed keynote speeches (whether his performance has improved since then is an issue we cannot judge). The results of Niebuhr et al. clearly show that, despite being less charismatic than SJ, MZ still performs significantly better on many acoustic-prosodic dimensions of perceived speaker charisma than the normal American English speaker who is, for example, asked to produce read monologues in a phonetic speech production experiment.[43]

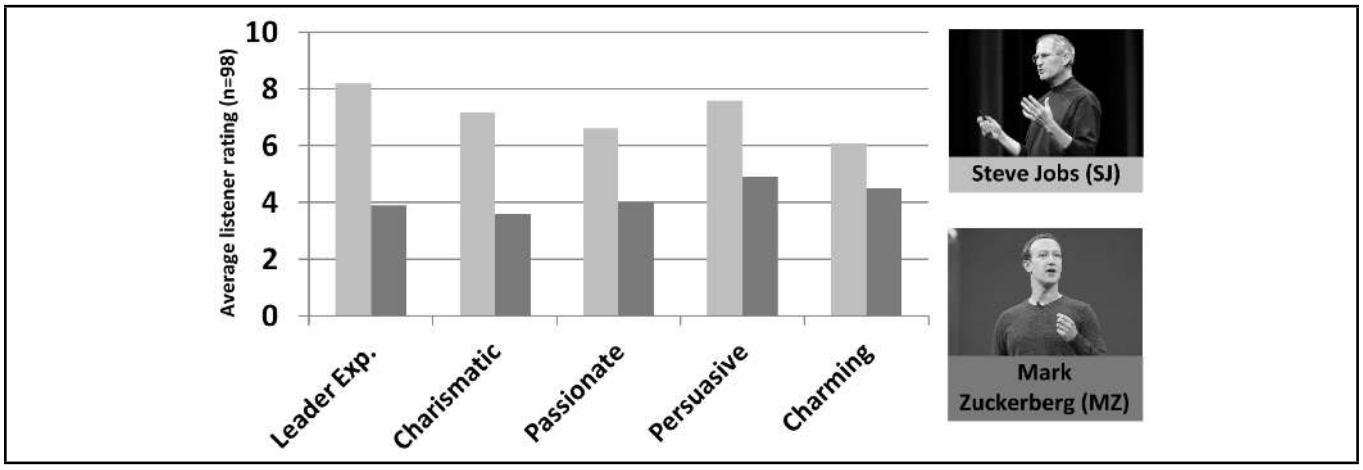

Figure 1. Results of the listener ratings ($n = 98$) on SJ's and MZ's charisma-related skills and personality traits. Photographs were taken by Ben Standfield (SJ at WWDC07) and by Anthony Quintano (MZ at F8 2018) and are used here in edited forms under flickr CC BY2.0 or CC BY-SA 2.0 licenses.

tity and the content of the speech were removed in the perception experiments of Niebuhr et al.[45] through the low-pass filtering, and since the visual channel was not available either (cp. Fox Cabane,[17] for charisma effects of attire, visual attractiveness, and body language), a significant part of the different charismatic impact of the two speakers must be contained in the speech signal. In fact, the two quotes of Sutter and Gruener above explicitly referred to the speech and voice of SJ and MZ, and such references were also made in many other similar comparative statements.[20,63] The results in Fig. 1 further corroborate the conclusions of previous studies that prosody is probably the most influential feature in charisma perception (see Niebuhr et al. for a summary).[44] Interestingly, these previous studies all used original hi-fi speech data, whereas the results in Fig. 1 were obtained solely with strongly reduced, low-pass filtered speech data. This suggested that there was a lot more to discover particularly in the variability and dynamics of prosodic parameters that is relevant for speaker charisma.

Taken all together, SJ and MZ represent two persons who are consistently judged to be very different in perceived charisma both by public actors like journalists and bloggers, and by naïve listeners in formal perception experiments. These judgment differences can clearly be related to the acoustic features in their speech signals. This certainly includes prosody, but, as the present study shows, probably extends beyond prosody to include the segmental domain as well.

SJ and MZ and the speech materials that are available for them allow keeping a number of things constant that could have otherwise become confounders in the analysis. SJ and MZ are male adults and were — at the times when their speeches were given — within an age range that is free from biologically-induced phonetic changes.[56] In addition, SJ and MZ both speak the same language (but different regional dialects, which is irrelevant, however, to the results of the present study, see section 4.2) and represent the same culture. Moreover, they are speakers for whom sufficient hi-fi speech material is available for acoustic analysis. This material was recorded in the same real-life communication situation (keynote speeches in large lecture halls with TV transmission in front of hundreds of listeners), it follows the same typical structure of a product presentation, and it comes from the same business area, i.e. digital, entertainment, and information technology.

The resilient differences in perceived charisma combined with the extensive control of many known individual and contextual factors influencing the production and perception of charismatic speech make SJ and MZ — or their keynote speeches, respectively — an ideal case study for the comparative acoustic analysis of speaker charisma. Nevertheless, note that the data we collect and analyse here still represent "field data". Thus, results obtained from this case study have a high degree of ecological validity and a high potential for generalization, but they will also need to be double-checked and further refined in more controlled lab-speech settings in which individual parameters can be singled out and separately varied or manipulated.

1.3. Consonant Realizations of Jobs and Zuckerberg

Based on the case-study arguments given above, Niebuhr et al. compared the consonant realizations of SJ and MZ in an acoustic analysis of keynote speech samples from the two speakers.[45] These samples were the same from which also the low-pass filtered stimuli were extracted for the perception experiment summarized in Fig. 1.

In particular, Niebuhr et al.[45] focused on SJ's and MZ's stop consonants, for which there are three pairs of phonological contrasts in American English, /p,t,k/ and /b,d,g/.[32] Although only a few hundred (out of several thousand) stop consonant realizations were acoustically analysed (in manual spectrogram and waveform measurements, carried out by Jana Thumm in the course of a funded EU young-researcher mobility program[4]), the results were clear.[45] Three examples of the overall results, all of which came out statistically significant, are shown in Figs. 2(a)–(c). The first two examples referred to the distinctiveness in the phonetic implementation of the phonological voiceless-voiced contrast. Compared to the keynote sample of SJ, MZ's sample was characterized by smaller differences in the closure duration of phonologically voiceless and voiced stops. This applied to all three places of articulation, but most strongly to the bilabial one. MZ's /p/ closures were realized on average only 17% longer than his /b/ closures, which was a relatively small percentage given that SJ's /p/ closures were on average almost 40% longer than his /b/ closures. For /t/ vs. /d/ and /k/ vs. /g/, SJ's closure durations

[4] See Niebuhr et al. for further methodological details of the acoustic analysis.[45]

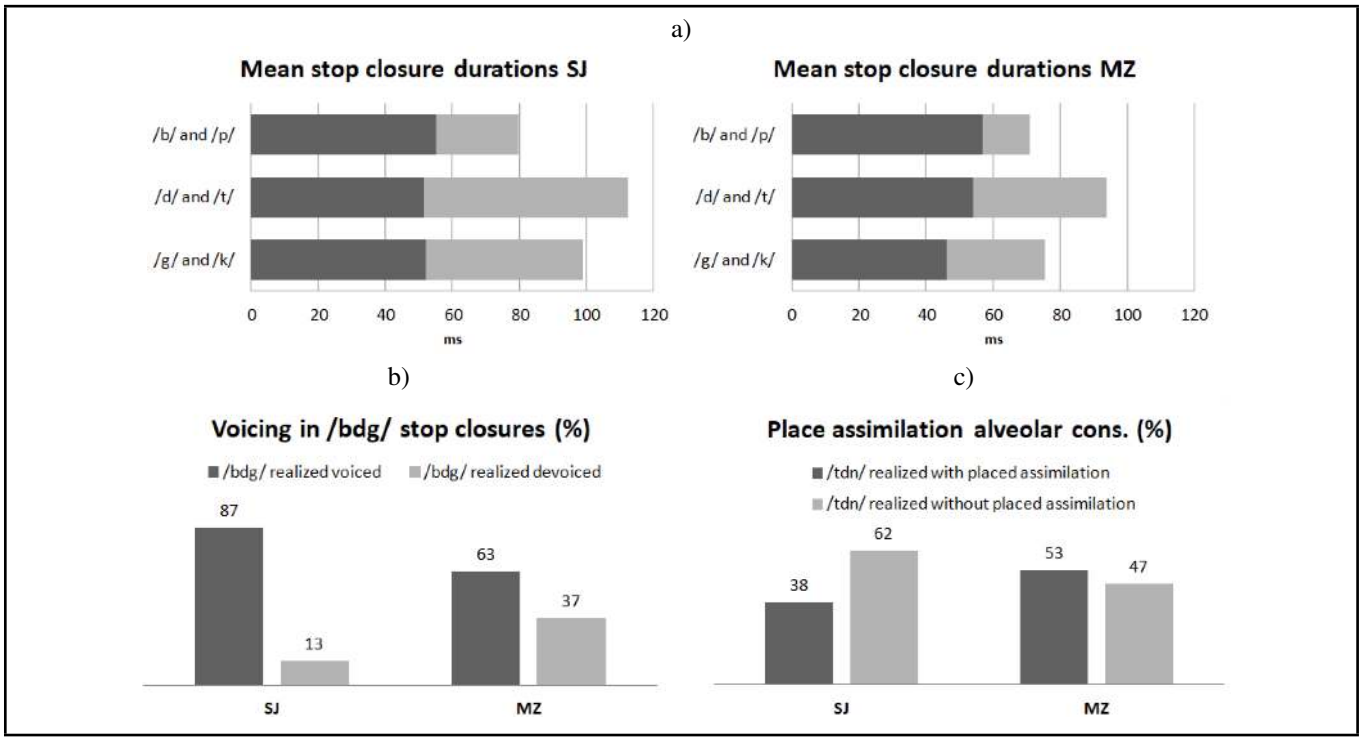

Figure 2. Summary of selected results from the acoustic analysis of SJ's and MZ's consonant realizations; (a) shows the mean closure durations of /b,d,g/ (dark bars) and the additional increase in closure durations for /p,t,k/ (light bars); (b) shows the percentages of /b,d,g/ realized mainly with (dark) or without (light) phonetic voicing during stop closure; (c) shows the percentages of /t,d,n/ realized with (dark) or without (light) post-lexical assimilation to a labial or velar place of articulation.

differed by almost a factor of two, which was about one-third more than those of MZ.

As regards the voicing distinction of the stop closures, results showed the phonologically voiced stops /b,d,g/ of SJ were also realized mostly with phonetic voicing during the stop closure. Only about 13% of SJ's /b,d,g/ tokens showed no consistent voicing during the stop closure, and were hence realized phonetically more similar to their phonologically voiceless /p,t,k/ counterparts. For MZ, this number was almost three times higher, i.e. 37%, see Fig. 2(b).

The third example in Fig. 2(c) concerns the degree to which SJ's and MZ's alveolar stop consonants /t/ and /d/ as well as the alveolar nasal /n/ were assimilated to other either labial or velar places of articulation. The assimilation analysis only included /t,d,n/ tokens in content words or content word sequences at the offset of lexically stressed syllables followed by labial or velar syllable onsets. That is, all reduction-prone "weak forms" (i.e. function words like articles, prepositions, auxiliary verbs etc., see Kohler) were excluded from the analysis.[30] Under these circumstances, the frequency of place assimilation of alveolar consonants was about 50% higher for MZ than for SJ. While a clear minority of about 38% of all analysed /t, d, n/ consonants were realized at other than alveolar places of articulation in SJ's keynote sample, a majority of 53% of /t, d, n/ underwent such an assimilation in MZ's keynote sample.

In summary, for multiple parameters that function as acoustic-cues to the phonological feature "voice" in American English (see Stevens, and Jiang et al.), MZ's speech showed a significantly lower level of phonetic distinctiveness in realizing the contrast between voiceless /p,t,k/ and voiced /b,d,g/.[27,60] In addition, the analysis reveals a higher level of speech reduction in terms of more frequent assimilations of alveolar consonants to either labial or velar places of articulation. Altogether, this means that the pronunciation of consonants (of context words) was significantly less clear in MZ's keynote sample than in SJ's keynote sample. Given that MZ is also perceived to be a less charismatic speaker than SJ, the results of the acoustic consonant analysis were consistent with the claim of rhetoric that a clear articulation supports the perception of speaker charisma.

1.4. Aim and Hypothesis

The findings outlined in section 1.3 lend initial support to the rhetorical claim that clearer speech is more charismatic speech. However, additional steps are necessary to further substantiate this claim. The second step after the first one made by Niebuhr et al. in section 1.3 was to extend this research to other types and classes of speech sounds than stop consonants.[45] On this basis, the third step was then to cross-check the acoustic findings in speech perception, i.e. by manipulating systematically the level of clarity of a speaker's speech across stimuli and then test whether these stimulus changes affect the perceived charismatic impact of a speaker.

The present paper dealt with the second step. That is, using the same dataset as in Niebuhr et al., our study extended the scope of the charisma-related acoustic sound-segment analysis from consonants to the other major class of sounds segments in speech: vowels.[45]

It was found many times (also across languages) that clear speech includes "an expansion of the vowel space."[58] Therefore, our research question was whether there is a systematic difference in the F1-F2 vowel spaces of SJ and MZ. More specifically, we tested the hypothesis that F1-F2 vowel space of the more charismatic speaker SJ is significantly larger than that of the less charismatic speaker MZ.

We did not include the third formant, F3, in the analysis for three reasons. First, F3 does not belong to the standard out-

put of the analysis software we use for this study. Separate measurements of F3 frequencies and their later addition to the dataset would have made our procedure inhomogeneous. Second, although F3 is to a small degree involved in acoustically separating American English vowels (in particular /ɝ/, /ɔ/, and /ʊ/, see Ghorshi et al.[18]), it is mainly studied for its critical role in the production and perception of rhoticity and variants of /r/ consonants in English, which is not the subject of the present study.[23] Hillenbrand and Gayvert (1987:3) supported the minor role of F3 in acoustically distinguishing American English vowels in their conclusion from the results of a large-scale automatic vowel classification: "The performance is reasonably good, and you can do everything by measuring just two parameters — F1 and F2."[24] Third, building on the second point, our aim was not to provide a fine-grained description, separation, and classification of American English vowels, but to find out whether SJ speaks clearer and, thus, uses a larger vowel space than MZ. To that end, it was sufficient to focus on F1 and F2 (which was also what was done by e.g., Smilijanic and Bradlow[58]), assuming that what applies to these two formants equally applies to F3.

2. METHOD

2.1. Speech Material

The analysed keynote speeches of SJ and MZ were obtained from YouTube. For Steve Jobs, we used two of his most well-known and influential speeches: the presentation of the iPhone 4 in 2010 and the presentation of the iPad 2 in 2011. Each presentation included the following sections that occurred in the same order in both speeches:

(1) Introduction: Welcoming. What has happened since the last presentation? What kinds of problems arose with products and how have they been solved? What updates are available?

(2) Main part I: Explanation of the company's development and current market position as well as the success and significance of the previous product(s); advantages over competitors.

(3) Main part II: Presentation of the new product. Its main new features and innovations are demonstrated, their advantages for the user are emphasized, sometimes in comparison to products of competitors.

(4) Main part III: Presentation and demonstration of further related innovations (e.g., apps); further information is provided on availability, price, and shipping of the new product; accessories for the new product are shown.

(5) Summary and acknowledgments.

Sections (2) and (3) are those that we extracted our speech data from. The speech in these middle sections was most consistent and free from familiarization artifacts or stylistic differences due to opening and closing addresses. Section (2) represents what we refer to as investor-oriented speech. Section (3) is defined as customer-oriented speech. About 11 minutes of speech were extracted from each section, 5–6 minutes from the iPhone 4 presentation, and another 5–6 minutes from the iPad 2 presentation. The start and end points of the extracted section were timed such that they coincided with prosodic-phrase boundaries of syntactically complete sentences. The extracted speech sections were to form a coherent stretch of speech with as few technical pauses, breaks, and audience-related interruptions (e.g., murmurs or applause) as possible. Besides that, the start point of the extraction was chosen randomly. The distinction between sections (2) and (3) was made as previous studies found prosodic differences between SJ and MZ addressing customers and investors.[43]

MZ's speech samples were extracted from his keynotes at Facebook's "F8" events.[54] F8 is Facebook's annual conference. It is meant to be a forum for highlighting milestones, advertising new features, and announcing the company's future plans and growth strategies. That is, MZ's F8 keynotes had the typical make-up of a product presentation and were hence similarly structured as those of SJ's. On this basis, we identified customer-oriented and investor-oriented sections on MZ's keynotes that met the same criteria as the sections (2) and (3) in SJ's keynotes. Then, following the same procedure as for SJ, we extracted 5–6 minutes from these sections in two successive F8 keynotes of MZ, the one from 2014 and the one from 2015.

In total, the present analysis is based on about 45 minutes of speech, i.e. 22 minutes from each of the two speakers, 11 minutes of investor-oriented, and 11 minutes of customer-oriented speech. The speech material was comprehensively annotated — in Praat TextGrid files[4] — at the levels of prosodic phrases, words, syllables, and individual sound segments. The data include 692 prosodic phrases of SJ and 532 prosodic phrases of MZ. For the annotation at the segmental level, we used the forced-alignment software DARLA, as this software is specifically trained for American English vowel phonemes.[51] Both sound and TextGrid files are available upon request under the following link: `10.5281/zenodo.1187140`.

After applying two data filters to the Praat TextGrid files — a grammatical filter and a segmental-context filter (see section 2.2) — a total of 1,990 vowels were analysed; 53% or 1,048 from SJ and 47% or 942 from MZ. In terms of the audience addressed, 46% of the analysed vowels were from the customer-oriented speech of the two speakers, and 54% from the investor-oriented speech. Note with respect to statistical processing and generalization that all vowel samples included at least 50 vowels per speaker and about half of the samples consisted of between 104 and 193 vowels. Broken down into the two sub-samples of customer- and investor-oriented speech, this meant that no statistically relevant sub-sample was smaller than 25 vowels per speaker; for most vowels, the sample size varied from 40 to 60 tokens. We considered these sample sizes sufficiently representative and reliable, not least because they were in the same order of magnitude as in other phonetic studies whose analysed phenomena range from dialectal vowel differences to coarticulation and the "segmental anchoring" of pitch-accent F0 peaks.[9,10,19,31]

2.2. Acoustic Analysis

Our analysis included the 10 American English vowel phonemes. They are shown in Table 1 together with example words and their representations in ARPA and the International Phonetic Alphabet (IPA). Table 1 also specifies the sample sizes per speaker and audience condition. We used DARLA for automatic vowel extraction. Form DARLA's *csv* output, vowel durations as well as F1 and F2 formant frequencies were

obtained.

The formant frequencies were measured at the temporal midpoint of each vowel and normalized using the Lobanov method.[16] Then, they were rescaled to Hz using the vowel package in R studio.[29] We took DARLA's csv output as it was and refrained from conducting manual checks or corrections for three reasons. First, DARLA is very effective in automatically filtering out noisy/ambiguous data based on a selective *in-dubio-contra-reo* principle. Second, comparative tests of forced-alignment/segmental-labelling software tools with American English speech material show that DARLA performs very well and "provides a clean, well-separated vowel space."[47] Third, we used a three-standard-deviation cut-off criterion per vowel phoneme in order to remove all potential remaining measurement errors from the DARLA output.

Regarding the two data filters mentioned in section 2.1, the first one excluded all vowels in grammatical (function) words, i.e. "weak forms" (see Kohler), from the analysis.[30] Thus, we focused on content words only, as the difference between content and function words is known to have a considerable effect on vowel reduction, more so than, for example, sentence accent.[67] Since we were dealing with field data, we could not extend control over the words any further. For example, we could not control the frequency of occurrence of the words whose target vowels we analysed. Studies like that of Wright showed that the frequency of a word is inversely related to how clearly the vowels inside this word are realized.[70] The more frequent a word is, the more reduced its vowels tend to be realized. However, this factor only manifested itself as noise in our data. A significant bias of this factor would require SJ to systematically use more/less frequent words in his keynotes than MZ. There were no indications for such a systematic difference in our keynote excerpts. Both SJ and MZ use a product- and company-specific inventory of less frequent words (e.g., technical terms) in combination with very frequent words with which they explain decisive pieces of information in a common language to customers, investors, and the general public. Nevertheless, we included the content words that contained the analysed vowels as a random factor in our statistical model. One thing we checked in this connection with respect to using field data was that no relevant content word and analysed target vowel was overlaid with environmental noises such as applause, laughter, camera clicking, etc.

The second applied data filter addressed the segmental context and excluded all vowel tokens adjacent to nasal consonants from the analysis. Nasals cause a co-articulatory nasalization of an adjacent vowel, and this changes the vowel's formant pattern and makes automatic formant extraction a very error-prone task.[9] Specifically, formants of nasalized vowels are more strongly damped than those of vowels in other co-articulation contexts. Damping increases the bandwidth of a formant, decreases its amplitude, and hence impedes the clear acoustic separation of adjacent formants like F1 and F2, especially in back vowels like [u], [ʊ], and [ɔ].[28] That the coupling of the nasal resonator cavities lowers formant frequencies and brings them closer together adds to the problem of acoustic formant separation.[14] Moreover, the coupling of the nasal resonator cavities introduces strong nasal formants at frequencies that can easily be mistaken for F1 and F2 of the vowel (e.g., at 250–300 Hz and 800–1,000 Hz). Simultaneously, the nasal sinuses represent side-tubes that branch off the nasal tube and create anti-formants in the vowel spectrum (Johnson, 2012).

Unlike other co-articulatory effects, nasal co-articulation extends very far into the adjacent vowel and can indeed affect the entire vowel. For example, Flege found in a speech-production experiment that on average 36% of the vowels of his American English speakers were realized fully nasalized in the context of an adjacent nasal consonant.[15] One-third of the speakers even realized 67–92% of all nasal-adjacent vowels fully nasalized. Besides the problems that nasal co-articulation causes for formant measurements, this strong co-articulatory effect was another reason why the segmental-context filter of the present study focused on nasals in particular. All other types of co-articulatory effects that are induced by adjacent segments in our target vowels are minimized by taking our formant measurements at the temporal midpoints of the vowels. It was repeatedly found and stressed by Reetz and Jogman[52] that formant frequencies at the vowel midpoint "provide the 'purest' representation of a vowel" in the sense that vowel midpoints are "most stable" against any contextual influences (see Larson and Hamlet, with reference to Picket).[33,49]

If small coarticulatory influences of neighbouring segments on the formant measurements at vowel midpoints remained in our data, then these influences only introduced some statistical noise to the data (as in the case of the word frequency above). We analysed about 1,000 vowels per speaker, and each vowel condition was represented by about 40–80 tokens (see Table 1). Sample sizes of this order of magnitude inherently cover a wide range of segmental contexts. This homogenized the compared vowel samples and, across the individual comparisons, turned potential effects of the segmental context into a random factor. That is, with respect to the tested hypothesis, neither speaker can consistently benefit from a segmental context effect.

Nevertheless, note that we also tested applying more restrictive filters to our dataset. These filters additionally excluded, for example, all those vowels that occurred in lexically unstressed and secondary stressed syllables, at the ends of prosodic phrases and/or in the segmental context of other vowels, liquids and semivowels. However, firstly, these additional prosodic and segmental filtering criteria substantially reduced our sub-sample sizes and, secondly, as we will discuss and illustrate in section 4.1, filtering out these additional contextual variations turned out to have no relevant effect on the results pattern with respect to our research question. Therefore, we decided to present the results of the largest dataset to which only the most essential grammatical and segmental-context filters had been applied.

3. RESULTS

The statistical tests are based on Generalized Linear Mixed Models (GLMMs). A separate test is run for each of the two measured formants, F1 and F2. Thus, formant (F1 or F2) represent the dependent variable. The fixed variables are Vowel (10 conditions), Speaker (2 conditions), and Audience (2 conditions). Target word is included as a random factor.

The overall results are similar for F1 and F2. We found a significant main effect only for the fixed factor Vowel (F1: F[9,1969]=739.9, $p < 0.001$, $\eta_p^2 = .776$; F2: F[9,1969] = 755.2, $p < 0.001$, $\eta_p^2 = .778$). The other two fixed factors, Speaker and Audience, have no separate significant main effects but yield significant two-way interactions with each other

Table 1. The analysed set of American English vowels, their ARPA labels (used in the Darla annotation) and symbols in the International Phonetic Alphabet (IPA). In addition, example words are given and sample sizes are specified (with the right number referring to the customer-oriented and the left to the investor-oriented sub-sample).

ARPA	AA	AE	AH	AO	EH	ER	IH	IY	UH	UW
IPA	ɑ	æ	ʌ	ɔ	ɛ	ɝ	ɪ	i	ʊ	u
Example	LOT	TRAP	STRUT	THOGHT	DRESS	NURSE	KIT	FLEECE	FOOT	GOOSE
Example	bot	bat	butt	bought	bet	bird	bit	beat	book	boot
Tokens SJ	33/41	64/69	25/33	41/40	48/80	79/82	39/65	88/89	27/36	29/40
Tokens MZ	33/35	77/74	25/30	37/61	39/45	48/63	32/40	96/97	26/25	26/33

(F1: F[1,1969] = 30.4, $p < 0.001$, $\eta_p^2 = .016$; F2: F[1,1969] = 7.4, $p = 0.007$, $\eta_p^2 = .004$) as well as with the fixed factor Vowel, i.e. Vowel*Speaker (F1: F[9,1969] = 7.7, $p < 0.001$, $\eta_p^2 = .035$; F2: F[9,1969] = 10.2, $p < 0.001$, $\eta_p^2 = .045$) and Vowel*Audience (F1: F[9,1969] = 2.3, $p = 0.013$, $\eta_p^2 = .011$; F2: F[9,1969] = 3.1, $p = 0.001$, $\eta_p^2 = .014$). The three-way interactions are not significant.

The separate significant main effect of Vowel means that the formant frequencies of the 10 vowel phonemes were distinct from one another. There are only a few exceptions, i.e. /ʌ/ vs. /ɛ/, /ɝ/ vs. /ʊ/, and /i/ vs. /u/ in the case of F1, and /æ/ vs. /ɛ/ and /ʌ/ vs. /ɝ/ in the case of F2. Both the distinctions and the exceptions to these distinctions fit in well with the acoustic F1 and F2 characteristics of American English vowel phonemes and, thus, may be seen as indirect evidence for the validity and reliability of the automatic formant-measurement procedure applied here.[10,19,24,48] On this basis, the significant interactions are looked at in detail by conducting post-hoc t-tests. They compare the formant measurements of each vowel phoneme between the two conditions of Speaker (section 3.1) and Audience (section 3.2). We use the conservative Bonferroni correction to compensate for multiple testing. Independent-sample and dependent-sample t-tests are conducted for Speaker- and Audience-related comparisons, respectively. Only significant outcomes are reported below, with grand means and p-levels in brackets.

3.1. Speaker

The overview displays of SJ's and MZ's customer-oriented and investor-oriented vowel spaces in Figs. 3(a)–(b) clearly show that SJ used a larger vowel space than MZ. From the perspective of MZ, SJ's vowel spaces are 6,680 Hz^2 or about 14% larger in the customer-oriented speech condition, and 17,973 Hz^2 or almost 50% larger in the investor-oriented speech condition. In total, i.e. across the two Audience conditions, the vowel space produced by SJ was about 12,500 Hz^2 or 33% larger than that of MZ. The vowel-space differences mainly concern the horizontal dimension, i.e. the frontness-backness dimension and its second formant F2. However, some vowel landmarks along the vertical dimension, i.e. the vowel height dimension and its first formant F1, are involved as well (see footnote 2). In addition, significant formant-frequency differences concern both phonologically long and phonologically short vowels.

Viewing and analysing both panels of Fig. 3(a)–(b) together shows that the vertical F1 difference in vowel space is primarily caused by three vowels and included the full range from high to low vowel qualities. Thus, SJ's vowel space extends beyond that of MZ in both vertical directions. High and mid vowels are produced higher, low vowels are produced lower. Specifically, SJ produced lower F1 values than MZ (i.e. higher vowel qualities) for /ʊ/ (333 Hz vs. 355 Hz, $p = 0.008$) and /ɔ/ (355 Hz vs. 390 Hz, $p = 0.003$), and higher F1 values (i.e. a lower vowel quality) for /æ/ (486 Hz vs. 458 Hz, $p < 0.001$).

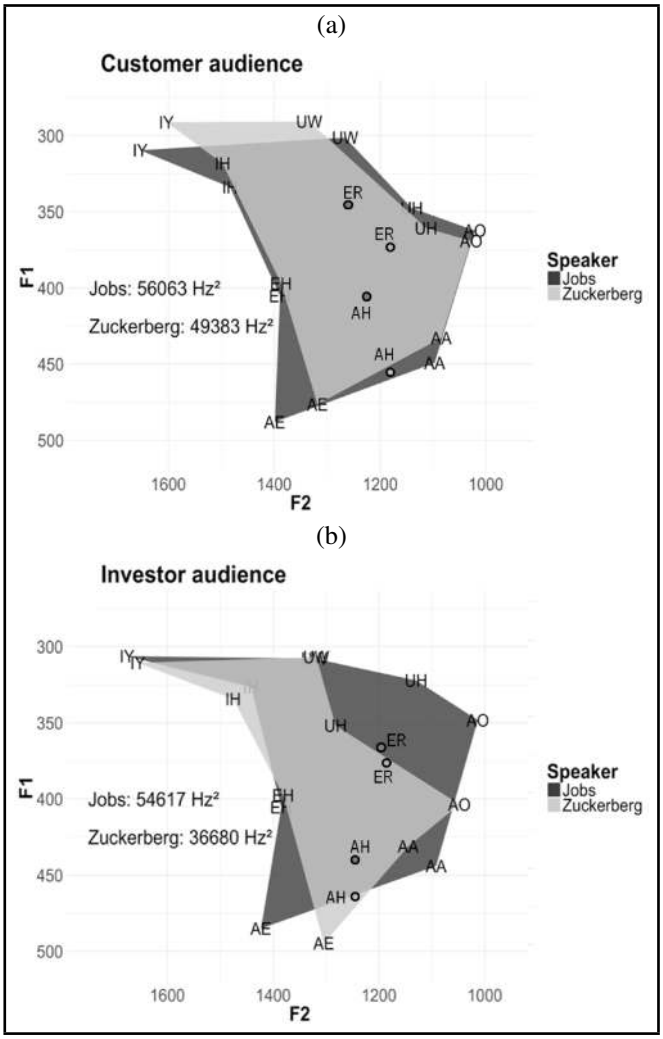

Figure 3. Vowel spaces (Hz2) in the customer-oriented and investor-oriented keynote sections of (a) SJ and (b) MZ; $n = 1{,}990$. Vowels are represented by their ARPA labels, see Table 1.

The same three vowels, plus the vowel /i/, are also primarily responsible for the horizontal F2 difference in vowel space. Again, vowels from both ends of the vowel space, i.e. back and front vowels, are involved and SJ's vowel space is expanded in both directions compared to that of MZ. SJ produced higher F2 values (i.e. more front vowel qualities) than MZ for /i/ (1,665 Hz vs. 1,630 Hz, $p = 0.01$) and /æ/ (1,412 Hz vs. 1,312 Hz, $p < 0.001$), and lower F2 values (i.e. more back vowel qualities) than MZ for /ʊ/ (1,134 Hz vs. 1,203 Hz, $p = 0.007$) and /ɔ/ (965 Hz vs. 1,094 Hz, $p = 0.03$).

3.2. Audience

Audience is a significant factor as well. Both SJ and MZ expand their vowel spaces when addressing customers in their keynotes; however, they do so in different ways and to different degrees, cp. panels (a) and (b) of Figure 3. In the case of SJ, for example, we found higher F1 values (i.e. lower vowel qualities) for /ʊ/ (348 Hz vs. 322 Hz, $p = 0.033$), /ɝ/ (353 Hz vs. 335 Hz, $p = 0.004$), and /ɪ/ (449 Hz vs. 429 Hz, $p = 0.033$) in the customer-oriented speech sample. Regarding F2, SJ produced higher values (i.e. a more front vowel quality) for /ɪ/ (1,486 Hz vs. 1,408 Hz, $p = 0.035$) and lower values (i.e. a more back vowel quality) for /u/ (1,165 Hz vs. 1,271 Hz, $p = 0.011$) in his customer-oriented speech sample. MZ's customer-oriented vowels showed lower F1 values (i.e. higher qualities) for /i/ (291 Hz vs. 311 Hz, $p < 0.001$), /ɪ/ (318 Hz vs. 334 Hz, $p = 0.003$), /u/ (291 Hz vs. 307 Hz, $p < 0.001$), and /ɔ/ (369 Hz vs. 404 Hz, $p = 0.035$). The F2 values of MZ decreased (i.e. towards more back vowel qualities) in his customer-oriented speech for the central vowels /ʌ/ (1,202 Hz vs. 1,235 Hz, $p < 0.001$) and /ɝ/ (1,194 Hz vs. 1,238 Hz, $p = 0.033$) and the back vowels /ɑ/ (1,048 Hz vs. 1,145 Hz, $p = 0.002$) and /ʊ/ (1,113 Hz vs. 1,282 Hz, $p = 0.023$). So, unlike for SJ, the vowel-space expansion in the customer-oriented speech of MZ involves both vowel height (high vowels are still higher) and horizontal tongue position (back vowels are further back). In consequence, MZ's vowel space expands by about 25.7% (12,703 Hz^2) from investor- to customer-oriented speech, whereas that of SJ expands by only about 2.7% (1,446 Hz^2).

4. DISCUSSION

4.1. Summary and Conclusions

Research and practice (e.g., manuals) in rhetoric and leadership have stressed for a long time how important it is for a charismatic speaker to have (i) a durable and animated voice and (ii) a clear and crisp articulation. While previous phonetic studies provided a large body of supporting evidence for the prosody-related claim (i) and detailed what "durable" and "animated" means in terms of acoustic-prosodic parameters, the segment-related claim (ii) has remained largely unaddressed so far. The aim of the present study was to address this knowledge gap in a case-study approach by means of a contrastive acoustic-phonetic analysis of Steve Jobs and Mark Zuckerberg. Given that previous studies showed that a clear articulation is reflected in a larger vowel space (see, for example, Smiljanić & Bradlow[58]), and given that Steve Jobs proved to be a more charismatic speaker than Mark Zuckerberg in public opinion as well as in a formal perception experiment, we tested the hypothesis that keynote speech excerpts of Steve Jobs would be characterized by a larger vowel space than those of Mark Zuckerberg. Our analysis of the first and second formant frequencies of about 2,000 phonologically long and short vowels supports this hypothesis: The vowel space exploited by Steve Jobs was larger than that of Mark Zuckerberg in both his customer-oriented and his investor-oriented keynote excerpts. As a larger vowel space means at the same time greater acoustic differences between adjacent vowel phonemes, we can conclude from our findings that Steve Jobs' phonological vowel contrasts were more distinctive at the phonetic level than those of Mark Zuckerberg. In other words, Steve Jobs realized larger differences between neighbouring vowels and may on this basis be regarded as speaking more clearly than Mark Zuckerberg,[5] in particular as our significant F1 and F2 differences are above the just noticeable difference (JND) of 0.37 bark and hence perceptually relevant.[37]

This conclusion is further corroborated by the fact that it is consistent with a previous comparison of Jobs' and Zuckerberg's consonant realizations by Niebuhr et al.[45] The analysis of consonants showed, for the same keynote excerpts as in the present study, that Mark Zuckerberg's speech contains significantly more place assimilations of alveolar consonants than Steve Jobs' speech and is also less distinct with respect to the acoustic implementation of the phonological "voiced" vs. "voiceless" opposition in stop consonants (see Fig. 2(a)–(c)). Jobs' speech shows greater differences for this opposition in terms of closure duration and vocal-fold vibration during the closure than Zuckerberg's speech.

In addition, the present finding is consistent with the previous finding on prosody that Steve Jobs and Mark Zuckerberg show a different speech behaviour in the customer- and investor-oriented parts of their keynotes. Both speakers are found here to produce more distinct and peripheral vowel qualities — and thus a larger vowel space — when addressing customers than when addressing investors, Mark Zuckerberg even more so than Steve Jobs. This is probably because customers are the primary type of audience for product presentations. In the previous study on prosody, however, Mark Zuckerberg was found to perform worse when addressing customers than when addressing investors.[43] The present data on sound segments go in the opposite direction. Mark Zuckerberg's vowel space is larger, i.e. he performed better for customers, not worse. The contrast between the prosodic and segmental findings suggests that Mark Zuckerberg applies different charisma-related phonetic implementation strategies in the two audience conditions. It seems as if he focuses more on intelligibility (the segmental aspects of his speech) when addressing customers and more on expressiveness (the prosodic aspects of this speech) when addressing investors, see section 4.2. Compared to that, Steve Jobs is not only the overall better speaker, but perhaps also more homogeneous with respect to his charisma-related phonetic implementation strategies.

One of our reviewers asks us for the role of experience in giving charismatic public speeches and if we would just have to wait for some twenty years until Mark Zuckerberg automati-

[5]Note for this conclusion that acoustic distinctivity between vowel phonemes is not solely a matter of the size of the vowel space. It must also be taken into account how much the acoustic F1-F2 ellipses of the individual vowels in the vowel space overlap. It was reasonable for us to assume that a small vowel space causes a larger overlap of the vowel-specific F1-F2 ellipses. For this reason, we focused on analysing the size of the vowel space. However, based on the comment of one of our reviewers, we conducted an additional post-hoc analysis of the acoustic overlap of the F1-F2 vowel ellipses. The results of this additional analysis confirm our assumption that a smaller vowel space is closely correlated with a larger overlap of the individual vowels in terms of their acoustic F1-F2 ellipses. For example, while the realizations of Steve Jobs' 10 vowels overlap acoustically on average by 288.4 Hz^2 across both audience conditions, the average overlap for Mark Zuckerberg is 377.8 Hz^2. This is considerably and statistically significantly larger than for Steve Jobs ($t[19,19] = -1.73$, $p = 0.042$). In addition, while for Steve Jobs no F1-F2 ellipse overlaps more than 50% with that of another vowel in the customer and the investor condition, there are two F1-F2 ellipses in Mark Zuckerberg's speech that are virtually completely covered by those of other vowels. This concerns /ɑ/ that acoustically fully overlaps with the larger F1-F2 ellipse of /ʌ/; and, in the investor condition, it additionally concerns /ɪ/, whose F1-F2 ellipse is completely covered by those of /ɝ/ and /u/.

cally performs as well as Steve Jobs. To the best of our knowledge, there is no solid empirical data about effects of mere public-speaking experience on a speaker's charismatic impact. However, the first author teaches a 10-week MSc course on "Persuasive Communication and Negotiation" at the University of Southern Denmark, which is mandatory for all business-engineering students. The students in this course learn how to give charismatic business-idea presentations in front of potential investors (so-called "investor pitches"). On this basis, he has some years ago determined how much his course participants improved over the 10 weeks, as compared to a baseline group of students who did not take part in the course but were instead instructed to practice without any rhetorical training and supervision a short-investor pitch presentation to other fellow students once a week for 10 weeks. Both the test group and the baseline group consisted of 17 male and female students. An analysis was done of the two groups in terms of the acoustic-prosodic parameters of charismatic speech (e.g., pitch range and level, speaking rate, etc.). Results show that both groups improved over time, the baseline group by about 20%, the test group about three times as much, i.e. by about 60%; while the test-group's improvement was fairly linear over time, the baseline-group's improvement curve took a clearly asymptotic shape at about 4 weeks. So, it seems speakers do become more charismatic by public-speaking experience, but that this improvement mainly concerns early-stage public speakers. Thus, we would not expect that time and experience alone will automatically make Mark Zuckerberg someday perform as well on stage as Steve Jobs. Steve Jobs was known for refining and rehearsing his presentations "endlessly and fastidiously."[57] We assume that it is this kind of training with an explicit focus on aspects of rhetoric and charisma that is required for an outstanding charismatic-speaking performance.

Finally, note that the presented differences in the vowel-space sizes of Steve Jobs and Mark Zuckerberg are very unlikely to result from coarticulatory artefacts. It is true that, in order to get large sub-samples for each vowel, we applied only very basic grammatical and segmental context filters to our speech data and, thus, included quite a bit of phonetic variation in the formant measurements of each vowel. However, applying more restrictive context filters to our speech data yields results that are qualitatively identical to those in Figs. 3(a)–(b). Figures 4(a)–(b) show the differences in vowel space between Steve Jobs and Mark Zuckerberg that emerge when filters additionally excluded all those vowels that occurred in lexically unstressed and secondary stressed syllables, at the ends of prosodic phrases, and in the segmental context of other vowels, liquids and semivowels. These restrictive filters reduce the total sample from almost 2,000 to slightly less than 1,000 vowel tokens ($n = 943$) in total.

Yet, again the vowel spaces of Steve Jobs are larger than those of Mark Zuckerberg in both audience and customer conditions; and again, this difference is more strongly pronounced for investor-oriented than customer-oriented speech; and again, we see that both the F1 and F2 dimensions are involved in the vowel-space differences regarding vowel-specific and audience-specific conditions. In a nutshell, we can say that the main effects and interaction patterns that result from the data in Figs. 4(a)–(b) are statistically equivalent to those of the data in Figs. 3(a)–(b). The only differences lie in which vowels contribute to these main effects and interactions and

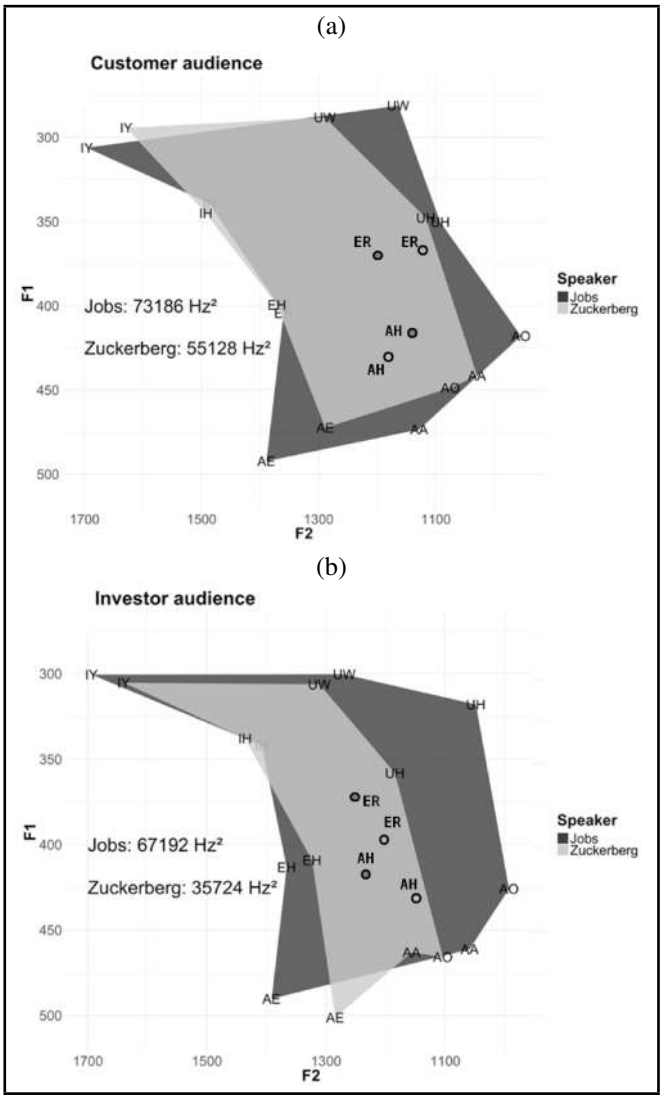

Figure 4. Vowel spaces (Hz2) in the customer-oriented and investor-oriented keynote sections of (a) SJ and (b) MZ, using a more restrictive phonetic-context filter than in the main analysis of Fig. 3; $n = 943$. Vowels are represented by their ARPA labels, see Table 1.

how strong these contributions are. Further aspects concerning the internal and external validity of the present results are discussed in the critical reflection below.

4.2. Critical Reflection and Outlook

The differences in vowel-space size between the two speakers can also not be an artefact of different dialects of American English. Although Steve Jobs and Mark Zuckerberg were born and raised in different regions of the US (San Francisco for Steve Jobs and White Plains for Mark Zuckerberg), the resulting differences in vocal production would not per se lend Steve Jobs a larger vocal space than Mark Zuckerberg.[10,19] In fact, we have not found many of the expectable dialectal differences between Steve Jobs and Mark Zuckerberg, which supports our auditory impression that neither Steve Jobs nor Mark Zuckerberg are strong dialect speakers. Also, a different vocal tract morphology (esp. length) or speaker age can be ruled out as explanations for our findings, either because they have opposite effects than those that were found here (e.g., a general vowel centralization for older speakers is expected, and we found the opposite for the older speaker Steve Jobs), or because these factors have comparable effects on all formants and would,

thus, only shift the vowel space as a whole along the F1-F2 plane, but not change its size.[56] The same applies to the factor of vocal effort, which equally raises formant frequencies (esp. F1) for all vowels.[34] Moreover, the major prosodic correlate of vocal effort, the intensity level (dB), is found to be higher for Mark Zuckerberg than for Steve Jobs.[43] So, if vocal effort were the explanation for our findings, then Mark Zuckerberg's vowel space would have been larger than Steve Jobs' vowel space, not smaller. Finally, although Steve Jobs speaks overall more slowly than Mark Zuckerberg (but still faster than the average male American English speaker, see Niebuhr et al.[42]), his vowels are only marginally — and statistically not significantly — shorter than those of Mark Zuckerberg (grand means: 115 ms vs. 108 ms); and Mark Zuckerberg's vowels are overall too long to entail a strong vowel-space compression. In addition, unlike the formant frequencies and Hz^2 values, Mark Zuckerberg's speaking rates (syl/s) did not differ between the two audience conditions.[43] For these reasons, the present findings can also not simply represent an epiphenomenon of an underlying speaking-rate difference.

All these points, in combination with the facts that our findings remain stable both for more and for less restrictive context filters (see Figs. 4(a)–(b) in section 4.1) as well as agree with previous measurements on consonants, lead us to the conclusion that Steve Jobs' speech was indeed characterized by a generally clearer and "crisper" articulation than Mark Zuckerberg's speech. Based on this empirical foundation, we can now, in a follow-up study, turn to the third step of the research agenda that we outlined in section 1.3. That is, we can proceed from acoustics to perception and test, if systematically manipulated variation in the level of articulatory precision and distinctiveness is actually linked to changes in perceived speaker charisma, and if so, how important this segmental articulation factor is for perceived charisma compared to other prosodic factors like pitch range and level (Berger et al., 2017), and whether precise and distinctive sound-segment articulation is more important for consonants than for vowels or vice versa.

There may also be interactions between prosodic and articulatory factors in charisma perception. For example, as was noted above, Steve Jobs' speaking rate is found to be slower than that of Mark Zuckerberg, but still faster than that of the average American English speaker. Thus, Steve Jobs manages to be both relatively fast and clear in his keynotes. Being fast is probably similarly beneficial for a speaker's charismatic impact as being clear (see Berger et al. for the positive correlation between speaking rate and perceived charisma).[3] However, there are natural physiological limits to the extent to which one can be a fast and clear speaker, and we assume that Mark Zuckerberg exceeds this limit in his keynote excerpts. His high speaking rate of on average more than 6 syl/s simply does not allow him (irrespective of whether or not he actually tried) to reach the same level of articulatory clarity and distinctiveness as Steve Jobs, who produces on average about 1 syllable less per second. It would be premature, though, to conclude on this basis and the two speakers' charisma differences that clear speech outweighs fast speech. However, the current data suggests testing this hypothesis.

In general, reconciling being fast and being clear is probably even harder for male than for female speakers (see Weirich et al.[68]), which is why follow-up studies will also have to look for gender differences in the articulation-charisma link. This also means extending the scope beyond Steve Jobs and Mark Zuckerberg, ideally to a level at which also cultural and situational differences as well as the language-specific phonologies and their acoustic leeway for less clear articulations come into play.

Regarding the leeway for less clear articulations, follow-up studies on charismatic speech will need to take into account that speech reduction is not per se bad. On the contrary, variation in the degree of speech reduction is to some degree context-determined (Clopper and Turnbull,[11] Ernestus and Smith[13]) and, on this basis, functional in speech communication.[41] For example, more or less strongly reduced sounds and syllables act as acoustic cues to turn-yielding and turn-holding, indicate lexical-stress positions and expressive meanings like irony, distinguish between given and new information, and help listeners predict the phonemic identity of subsequent sound segments. That is, speech communication *needs* constant variation in the degree of speech reduction, which, in turn, means that clear articulation in charismatic speaking cannot simply mean producing each and every word as specified in a pronunciation dictionary, although this is what the Effort Code and the H&H theory would probably assume (and what rhetoric manuals imply with sweeping imperatives like "speak clearly!"). In fact, the perception experiment of Niebuhr showed that such a constant overly-clear articulation can make a speaker sound arrogant and vain; attributes that could hardly be further away from those associated with a charismatic speaker but represent key characteristics of a hubristic (business) speaker, cf. Sundermeier.[44,62]

This raises highly relevant questions about what a clear articulation means in detail for charismatic speakers. Does a more charismatic speaker show the same variation in the degree of speech reduction as does a less charismatic speaker, but at a generally higher (clearer) level? Or, given that an increased variability (or the avoidance of any kind of monotony) is a key characteristic of charismatic speech (e.g., Strangert and Gustafson[61] and Hiroyuki and Rathcke[25]), does being a more charismatic speaker actually mean exceeding the variation in the degree of speech reduction of less charismatic speakers, i.e. to be even clearer when they are clear and to reduce even more when they are less clear? Since the speech-reduction level is functional in communication, an exceeded variation in the degree of speech reduction is also what one would assume, if charismatic speech means a greater acoustic distinctiveness of phonological contrasts. These questions are not the first on our research agenda, but they are questions whose answers will also contribute to elaborating the theoretical framework of charismatic speech and the roles and interconnections of expressiveness and intelligibility within this framework. For instance, as was briefly suggested in section 4.1, the fact that Mark Zuckerberg shows a better prosody-related charisma performance in his investor-oriented keynote sections and a better articulation-related charisma performance in this customer-oriented keynote sections could mean that his charisma effect was overall similar for both audience groups, but more expressiveness-based in the ears of investors and more intelligibility-based in the ears of customers. Such an idea could be tested by combining behavioural and physiological measurements in perception experiments. For example, in an eye-tracking setup, reactions times (e.g., based on phoneme-monitoring tasks) could measure intelligibility,

while the eye tracker itself measures expressiveness in the form of pupil dilation. EEG data could even be able to directly measure both intelligibility and expressiveness perception in different areas of the listener's brain.

Examining to what degree the prosodic and segmental characteristics of charismatic speech are related to increasing a speaker's expressiveness and intelligibility and determining on this basis how important these two features are for perceived charisma is probably one of the major tasks of the phonetic sciences in charisma research. It would also allow us someday to predict how charismatic speech has to be adapted to different environmental as well as second-language contexts.

ACKNOWLEDGEMENTS

The authors would like to thank their four anonymous reviewers for their insightful and constructive ideas and comments on earlier drafts of this manuscript. We are also greatly indebted to J. P. Arenas for his thorough handling of our manuscript during the peer-review process. Further thanks a due to Ferran Giones, Plinio Barbosa, Radek Skarnitzl, Jan Michalsky, Stephanie Berger, Jana Voße, Ocke-Schwen Bohn and all organizers and participants of the Interspeech special session on "voice attractiveness" (2017) for their inspiring and committed discussions on the concept and nature of charismatic speech and its phonetic ingredients. Finally, we owe special thanks to Donna Erickson for her careful proof-reading of the resubmitted paper and her great ideas about articulation-related follow-up studies on speaker charisma.

REFERENCES

[1] Antonakis, J., Bastardoz N., and Jacquart P. Charisma: An ill-defined and ill-measured gift, *Annual Review of Organizational Psychology and Organizational Behaviour*, **3**, 293–319, (2016). https://dx.doi.org/10.1146/annurev-orgpsych-041015-062305

[2] Awamleh R., and Gardner, W. L. Perceptions of leader charisma and effectiveness: The effects of vision content, delivery, and organizational performance, *The Leadership Quarterly*, **10** (3), 345–373, (1999). https://dx.doi.org/10.1016/s1048-9843(99)00022-3

[3] Berger, S., Niebuhr, O., and Peters, B. Winning over an audience — a perception-based analysis of prosodic features of charismatic speech, *Proc. 43rd Annual Conference of the German Acoustical Society (DAGA)*, Kiel, Germany, (2017).

[4] Boersma, P. Praat, a system for doing phonetics by computer, *Glot International*, **5**, 341–345, (2001). https://dx.doi.org/10.1097/aud.0b013e31821473f7

[5] Biadsy, F., Rosenberg A., Carlson R., Hirschberg J., and Strangert E. A cross-cultural comparison of American, Palestinian, and Swedish perception of charismatic speech, *Proc. 4th International Conference of Speech Prosody*, Campinas, Brazil, (2008).

[6] Bosker, H. R. The role of temporal amplitude modulations in the political arena: Hillary Clinton vs. Donald Trump, *Proc. 18th International Interspeech Conference*, Stockholm, Sweden, (2017). https://dx.doi.org/10.21437/interspeech.2017-142

[7] Camper Bull, R. *Moving from project management to project leadership: a practical guide to leading groups*, CRC, Boca Raton, (2010). https://dx.doi.org/10.1201/9781439826683

[8] Chen, A., Gussenhoven, C., and Rietveld, A. Language-specific uses of the Effort Code, *Proc. 1st International Conference of Speech Prosody*, Aix-en-Provence, France, (2002).

[9] Cho T., Kim D., and de Jong K. Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in English, *Journal of Phonetics*, **64**, 71–89, (2017). https://dx.doi.org/10.1016/j.wocn.2016.12.003

[10] Clopper C. G., Pisoni D. B., and de Jong, K. Acoustic characteristics of the vowel systems of six regional varieties of American English, *Journal of the Acoustical Society of America*, **118** (3), 1661–1676, (2005). https://dx.doi.org/10.1121/1.2000774

[11] Clopper, C. G., and Turnbull, R. Exploring variation in phonetic reduction: Linguistic, social, and cognitive factors, in F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler, and M. Zellers (eds), Rethinking Reduction: Interdisciplinary Perspectives on Conditions, *Mechanisms, and Domains for Phonetic Variation*, F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler, and M. Zellers (eds), de Gruyter, Berlin/Boston, 25–72, (2018). https://dx.doi.org/10.1515/9783110524178-002

[12] Eriksson, A., and Traunm´uller, H. Perception of vocal effort and distance from the speaker on the basis of vowel utterances, *Perception & Psychophysics*, **64** (1), 131–139, (2002). https://dx.doi.org/10.3758/bf03194562

[13] Ernestus, M., and Smith, R. Qualitative and quantitative aspects of phonetic variation in Dutch "eigenlijk", in F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler, and M. Zellers (eds), *Rethinking Reduction: Interdisciplinary Perspectives on Conditions, Mechanisms, and Domains for Phonetic Variation*, F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler, and M. Zellers (eds), de Gruyter, Berlin/Boston, 129–163, (2018). https://dx.doi.org/10.1515/9783110524178-005

[14] Feng, G., and Castelli, E. Some acoustic features of nasal and nasalized vowels. A target for vowel nasalization, *Journal of the Acoustical Society of America*, **99** (6), 3694–3706, (1996). https://dx.doi.org/10.1121/1.414967

[15] Flege, J. E. Anticipatory and carry-over nasal coarticulation in the speech of children and adults, *Journal of Speech and Hearing Research*, **2** (4), 525–536, (1988). https://dx.doi.org/10.1044/jshr.3104.525

[16] Flynn, N., and Foulkes, P. Comparing vowel formant normalization methods, *Proc. 17th International Congress of Phonetic Sciences*, Hong Kong, China, (2011).

[17] Fox Cabane, O. *The charisma myth: How anyone can master the art and science of personal magnetism*, Penguin, New York, (2012).

[18] Ghorshi, S. A., Vaseghi, S., and Yan, Q. Comparative analysis of formants of British, American and Australian accents, *Proc. 7th International Interspeech Conference*, Pittsburgh, USA, (2006).

[19] Grieve J., Speelman D., and Geeraerts D. A multivariate spatial analysis of vowel formants in American English, *Journal of Linguistic Geography*, **1** (1), 31–51, (2013). https://dx.doi.org/10.1017/jlg.2013.3

[20] Gruener W. *On stage*, Mark Zuckerberg is no Steve Jobs, Tom's Guide (2011). Retrieved from http://www.tomsguide.com/us/facebook-skype-video-calling-social-networking,news-11808.html, (Accessed December 28, 2018).

[21] Gussenhoven, C. Intonation and interpretation: phonetics and phonology, *Proc. 1st International Conference of Speech Prosody*, Aix-en-Provence, France, (2002).

[22] Gussenhoven C. *The phonology of tone and intonation*, CUP, Cambridge: (2004). https://dx.doi.org/10.1017/cbo9780511616983

[23] Heselwood, B., and Plug, L. The Role of F2 and F3 in the perception of rhoticity: evidence from listening experiments, *Proc. 17th International Congress of Phonetic Sciences*, Hong Kong, China, (2011).

[24] Hillenbrand, J., and Gayvert, R. NAIC Technical Series Report, *Speaker-Independent Vowel Classification Based on Fundamental and Formant Frequencies*, (1987). https://dx.doi.org/10.1121/1.2024478

[25] Hiroyuki, T., and Rathcke, T. Then, what is charisma? The role of audio-visual prosody in L1 and L2 political speeches, *Proc. Phonetik & Phonologie im deutschsprachigen Raum*, Munich, Germany, (2016).

[26] Holtgraves, T. M. Language as social action: social psychology and language use, *Lawrence Erlbaum*, Mahwah, (2001). https://dx.doi.org/10.4324/9781410601773

[27] Jiang, J., Chen, M., and Alwan, A. On the perception of voicing in syllable-initial plosives in noise, *Journal of the Acoustical Society of America*, **119** (2), 1092–1105, (2006). https://dx.doi.org/10.1121/1.2149841

[28] Johnson, K. *Acoustic and Auditory Phonetics*, Blackwell, Oxford, (2012). https://dx.doi.org/10.1159/000078663

[29] Kendall, T., and Thomas, E. R. Vowels: Vowel manipulation, normalization, and plotting in R, (2009). Retrieved from http://cran.r-project.org/web/packages/vowels/index.html, (Accessed August 14, 2018).

[30] Kohler, K. J. *Segmental reduction in connected speech in German: phonological facts and phonetic explanations, in Speech Production and Speech Modelling*, W.J. Hardcastle, and A. Marchal (eds), Kluwer Academic Publishers, Dordrecht, (1990), 69–92. https://dx.doi.org/10.1007/978-94-009-2037-8_4

[31] Ladd D. R., Faulkner D., Faulkner H., and Schepman A. Constant 'segmental anchoring' of F0 movements under changes in speech rate, *Journal of the Acoustical Society of America*, **106** (3), 1543–1554, (1999). https://dx.doi.org/10.1121/1.427151

[32] Ladefoged, P. *American English, The Handbook of the International Phonetic Association*, Cambridge University Press, Cambridge, 41–44, (1999). https://dx.doi.org/10.1017/s0952675700003894

[33] Larson, P. L., and Hamlet, S. L. Coarticulation effects on the nasalization of vowels using nasal/voice amplitude ratio instrumentation, *The Cleft Palate Journal*, **24** (4), 286–290, (1987).

[34] Liénard, J. S. and Di Benedetto, M. G. Effect of vocal effort on spectral properties of vowels, *Journal of the Acoustical Society of America*, **106** (1), 411–422, (1999). https://dx.doi.org/10.1121/1.428140

[35] Lindblom B. Explaining phonetic variation: A sketch of the H & H theory, in Speech Production and Perception, W. J. Hardcastle, A. Marchal (eds), Kluwer, Dordrecht, 403–439, (1990). https://dx.doi.org/10.1007/978-94-009-2037-8_16

[36] Lisker, L., and Abramson, A. S. A cross-language study of voicing in initial stops: acoustical measurements, *Word*, **20** (3), 384–422, (1964). https://dx.doi.org/10.1080/00437956.1964.11659830

[37] Liu C. and Kewley-Port, D. Vowel formant discrimination for high-fidelity speech, *Journal of the Acoustical Society of America*, **116** (2), 1224–1233, (2004). https://dx.doi.org/10.1121/1.1768958

[38] Louekari L. *Charismatic communication style in knowledge-intensive organizations*, MA thesis, School of Business, Aalto University, Finland, (2015).

[39] Miet, G., Gerrits, A., and Valière, J. C. Low-band extension of telephone band speech, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Edinburgh, Scotland, (2000). https://dx.doi.org/10.1109/icassp.2000.862116

[40] Mortensen, K. W. The laws of charisma: how to captivate, inspire, and influence for maximum success: how to captivate, inspire, and influence for maximum success, *Amacom*, New York, (2011).

[41] Niebuhr, O. Rich reduction: sound-segment residuals and the encoding of communicative functions along the hypo-hyper scale, *Proc. 7th Tutorial & Research Workshop on Experimental Linguistics*, St. Petersburg, Russia, (2016).

[42] Niebuhr, O., Voße, J., and Brem, A. What makes a charismatic speaker? A computer-based acoustic-prosodic analysis of Steve Jobs tone of voice, *Computers in Human Behaviour*, **64**, 366–382, (2016). https://dx.doi.org/10.1016/j.chb.2016.06.059

[43] Niebuhr, O., Brem, A., and Novák-Tót, E. Prosodic constructions of charisma in business speeches — A contrastive acoustic analysis of Steve Jobs and Mark Zuckerberg, *Proc. 8th International Conference of Speech Prosody*, Boston, USA, (2016).

[44] Niebuhr, O. Clear speech — mere speech? How segmental and prosodic speech reduction shape the impression that speakers create on listeners, *Proc. 18th Interspeech Conference*, Stockholm, Sweden, (2017). https://dx.doi.org/10.21437/interspeech.2017-28

[45] Niebuhr O., Thumm J., and Michalsky J. Shapes and timing in charismatic speech - Evidence from sounds and melodies, *Proc. 9th International Conference of Speech Prosody*, Poznan, Poland, (2018). https://dx.doi.org/10.21437/speechprosody.2018-118

[46] Novák-Tót, E., Niebuhr, O., and Chen, A. A gender bias in the acoustic-melodic features of charismatic speech? *Proc. 18th Interspeech Conference*, Stockholm, Sweden, (2017). https://dx.doi.org/10.21437/interspeech.2017-1349

[47] Olsen, R. M., Olsen, M. L., Stanley, J. A., Renwick, M., and Kretzschmar, W. A. Methods for transcription and forced alignment of a legacy speech corpus, *Proc. ASA Meetings on Acoustics*, **30** (1), 1–13, (2017). https://dx.doi.org/10.1121/2.0000559

[48] Peterson, G, and Barney, H. Control methods used in a study of the vowels, *Journal of the Acoustical Society of America*, **24** (2), 175–184, (1952). https://dx.doi.org/10.1121/1.1906875

[49] Pickett, J. M. *The Sounds of speech communication: A primer of acoustic phonetics and speech perception*, University Park Press, Baltimore, (1980).

[50] Ramus, F., Nespor, M., and Mehler, J. Correlates of Linguistic Rhythm in the Speech Signal, *Cognition*, **73**, 265–292, (1999). https://dx.doi.org/10.1016/s0010-0277(99)00058-x

[51] Reddy, S., and Stanford, J. Toward completely automated vowel extraction: introducing DARLA, *Linguistics Vanguard*, **1** (1), 15–28, (2015). https://dx.doi.org/10.1515/lingvan-2015-0002

[52] Reetz, H., and Jongman, A. *Phonetics: transcription, production, acoustics, and perception*, Blackwell, Oxford, (2009).

[53] Rosenberg A., and Hirschberg, J. Charisma perception from text and speech, *Speech Communication*, **51** (7), 640–655, (2009). https://dx.doi.org/10.1016/j.specom.2008.11.001

[54] Rusli, E. M. Facebook to hold developer conference on April 30, *The Wall Street Journal*, (2014). Retrieved from http://blogs.wsj.com/digits/2014/03/08/facebook-to-hold-f8-developer-conference-on-april-30/, (Accessed February 4, 2016).

[55] Samlowski, B., Kern, F., and Trouvain, J. Perception of Suspense in Live Football Commentaries from German and British Perspectives, *Proc. 9th International Conference of Speech Prosody*, Poznan, Poland, (2018). https://dx.doi.org/10.21437/speechprosody.2018-8

[56] Schötz S. *Perception, analysis, and synthesis of speaker age*, PhD thesis, Lund University, Sweden, (2006).

[57] Schlender, B. and Tetzeli, R. *Becoming Steve Jobs*, New York, Crown, (2015).

[58] Smiljanić R., and Bradlow, A. R. Production and perception of clear speech in Croatian and English, *Journal of the Acoustic Society of America*, **118** (3), 1677–1688, (2005). https://dx.doi.org/10.1121/1.2000788

[59] Sørensen, L. S. *How to grow an Apple: did Steve Jobs speak Apple to success?*, MA thesis, Aarlborg University, Denmark, (2013).

[60] Stevens, K. N. *Acoustic phonetics*, MIT Press, Cambridge, (1998).

[61] Strangert, E. and Gustafson, J. What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations, *Proc. 9th International Interspeech Conference*, Brisbane, Australia, (2008).

[62] Sundermeier, J. How does hubris influence the innovation activities of company founders? *Proc. Babson Conference Entrepreneurship Research Conference*, Bodø, Norway, (2016).

[63] Sutter, J. D. *When it comes to presentation, Mark Zuckerberg is no Steve Jobs*, (2011). Retrieved from http://edition.cnn.com, (Accessed December 28, 2018).

[64] Tobak, S. *Charisma: an obsolete leadership quality*, CBS News, (2012). Retrieved from http://www.cbsnews.com/news/charisma-an-obsolete-leadership-quality/, (Accessed December 28, 2018)

[65] Touati, P. Prosodic aspects of political rhetoric, *Proc. ESCA Workshop on Prosody*, Lund, Sweden, (1993).

[66] Traunmüller, H. and Eriksson, A. Acoustic effects of variation in vocal effort by men, women, and children, *Journal of the Acoustical Society of America*, **107** (6), 3438–3451, (2000). https://dx.doi.org/10.1121/1.429414

[67] van Bergem D. R. Acoustic vowel reduction as a function of sentence accent, word stress, and word class, *Speech Communication*, **12** (1), 1–23, (1993). https://dx.doi.org/10.1016/0167-6393(93)90015-d

[68] Weirich, M., Fuchs, S., Simpson, A., Winkler, R., and Perrier, P. Mumbling: macho or morphology?, *Journal of Speech, Language, and Hearing Research*, **59**, S1587–S1595, (2016). https://dx.doi.org/10.1044/2016_jslhr-s-15-0040

[69] Weninger, F., Krajewski, J., Batliner, A., and Schuller, B. The voice of leadership: Models and performances of automatic analysis in online speeches, *IEEE Transactions on Affective Computing*, **3** (4), 496–508, (2012). https://dx.doi.org/10.1109/t-affc.2012.15

[70] Wright, R. *Factors of lexical competition in vowel articulation, in Papers in Laboratory Phonology VI, Local J., Ogden R., Temple R. (eds)*, Cambridge University Press, Cambridge, 75–86, (2004). https://dx.doi.org/10.1017/cbo9780511486425.005

[71] Xu, M., Homae, F., Hashimoto, R., and Hagiwara, H. Acoustic cues for the recognition of self-voice and other-voice, *Frontiers in Psychology*, **4** (735), 1–7, (2013). https://dx.doi.org/10.3389/fpsyg.2013.00735