THE DYNAMIC EFFECT OF SPEAKING FAST ON SPEECH PROSODY

LAURI TAVI

ABSTRACT

Speaking fast causes several changes in speech prosody. In addition, it can be associated with a decrease in speech intelligibility. In this study, prosodic changes in fast speech were investigated using common prosodic measurements and syllabic prosody index (SPI), a novel prominence measure that combines f0, energy and duration features. Dynamic changes in long-term prosodic prominence were investigated using functional data analysis (FDA), in which the SPI is transformed into a functional form. The possibly decreasing effect of speaking fast on speech intelligibility was evaluated using automatic speech recognition. Phonetic analyses of syllabic units showed that speaking fast decreases duration, f0 and SPI, and increases articulation rate and proportional acoustic energy in the frequency range of 0–1 kHz. FDA supported the aforementioned results by revealing dynamically decreased overall prominence in fast speech. Furthermore, in comparison to regular speech, speech intelligibility was found to be significantly lower in fast speech: word error rate (WER) for regular speech was 0.27, whereas for fast speech it was 0.86.

Keywords: fast speech, prosody, prominence, functional data analysis, speech intelligibility

1. Introduction

It is well known that speech characteristics, such as prosodic features and articulatory gestures, change dynamically over time (Roettger et al., 2019; Niebuhr et al., 2011). The dynamic changes occur particularly in natural speech communication, in which speakers alter their speech both voluntarily and involuntarily. According to Lindblom's theory of Hyper and Hypoarticulation (H&H), for example, speakers intentionally adapt their speech according to conversational demands (Lindblom, 1990). In other words, speakers' articulatory effort can decrease (hypoarticulation) or increase (hyperarticulation) depending on how intelligible they believe their speech is for listeners. An example of an involuntary change in prosody is the Lombard effect, which causes speakers to increase their vocal effort in noisy conditions (Stanton et al., 1988; Patel & Schell, 2008). These changes include increases of pitch and duration of words, yielding improved speech audibility and intelligibility.

Some prosodic changes, such as decrease in word duration, can also decrease speech intelligibility (Mayo et al., 2012; Hazan & Markham, 2004). In addition to the

https://doi.org/10.14712/24646830.2022.28

fact that words may be less carefully articulated in fast speech (Janse, 2004), the timing patterns are different from those in regular speech tempo. When speaking fast, the duration of unstressed syllables is reduced more than that of stressed syllables, resulting in a more prominent prosodic pattern (Janse, 2004). The more prominent pattern in fast speech, however, probably does not improve intelligibility. In fact, increasing the speech rate artificially without changing prominence patterns has yielded speech that is more difficult to process compared to naturally fast speech (Janse et al., 2003; Janse, 2004).

Prosodic prominence refers to the relative emphasis of syllables, which can be acoustically measured as the variation of relative energy, duration and fundamental frequency (f0) (Greenberg et al., 2003; Tavi & Werner, 2020). Although the term "prominence" typically refers to relative changes, for example, between adjacent syllables, here the term is also used to describe the strength of emphasis between different speaking conditions. During fast speech, speakers might not be able to properly emphasize syllables due to the limited articulation and processing time, which could result in a decrease of overall prominence. However, to the author's knowledge, previous phonetic studies lack acoustically orientated analyses of exactly how fast speech impacts the dynamics of prosodic prominence in healthy speakers.

In order to establish the relationship between fast speech and prosodic prominence, two hypotheses were formulated and tested in this study: Speaking fast (1) decreases prosodic prominence and (2) impairs speech intelligibility. The possible decrease of prosodic prominence is examined focusing on different aspects of prosody, i.e., pitch, energy and durational characteristics. To avoid subjective listening tests, speech intelligibility was evaluated using the accuracy of automatic speech recognition (ASR) in terms of word error rate (WER). WER has been a common metric to evaluate speech intelligibility in an objective and comparable manner in numerous technology-oriented speech studies, such as in Voice Privacy Challenge (Tomashenko et al., 2021).

In addition to inspecting conventional statistics, this study utilizes functional data analysis (FDA) in order to reveal wide-scale dynamic differences in prosodic prominence between regular and fast speech. The focus is on the steadiness, or major shifts, of long-term prominence rather than on high-frequency prominence variation between adjacent syllables. FDA is a methodology which extends conventional statistics from discrete values to functions of time (Ramsay et al., 2009). One popular method in FDA has been functional principal component analysis (fPCA), in which eigenvalues are paired with eigenfunctions instead of eigenvectors as in traditional PCA. In previous phonetic studies, fPCA has been shown to be an effective method of capturing the dynamic nature of speech (Cronenberg et al., 2020; Gubian et al., 2010, 2011; Zellers et al., 2010; Gubian et al., 2015).

In this paper, Section 2 will describe the speech data and analysis techniques used in this study. The answers to the aforementioned hypotheses will be presented in Section 3 and discussed in Section 4.

2. Speech data and methods

2.1 The Chains corpus

Prosodic characteristics of fast speech are investigated using the Chains corpus. The Chains corpus was collected in 2005 in order to study challenges in speaker identification (Cummins et al., 2006). The corpus contains six different speaking conditions (i.e., retelling, synchronous imitation, repetitive synchronous, solo, fast, and whispered speech) from 36 (20 male and 16 female) speakers. The speakers read aloud four short fables (”Cinderella”, ”Rainbow text”, ”North Wind and the Sun”, and ”Members of the Body”) and 33 individual sentences. In this study, only the four fables produced with solo (hereafter referred to as “regular”) and fast speaking conditions were analysed.

2.2 Phonetic measurements

Phonetic analyses were carried out with Praat (Boersma and Weenink, 2020) and performed separately for the female and male speakers. Firstly, the readings of the four fables were automatically segmented into syllabic units using Vocal toolkit (Corretge, 2020), which adapts a script (De Jong and Wempe, 2009) to detect syllable nuclei. The term ”syllabic unit” is used here because automatic syllable markings are not perfectly aligned with linguistic syllables.

Secondly, a total of five acoustic-phonetic features were analysed from the syllabic units: articulation rate, duration, relative energy proportion below 1 kHz in a frequency range of 0–4 kHz, median f0 and syllabic prosody index (SPI). The SPI (Tavi and Werner, 2020) measures prosodic prominence in syllables by combining their pitch, duration and energy proportion below 1 kHz into one feature. SPI is formulated as

$$SPI = \frac{Pitch_{median} \times \sqrt{Duration}}{\sqrt{Energy_{below1kHz}}}/10.$$

The higher the SPI, the higher the prominence in a syllabic unit. In pitch analysis, the ceiling and the floor values were set to 120 and to 500 Hz for female speakers and to 70 and to 400 Hz for male speakers. The relative energy proportion was calculated by dividing the overall energy in the frequency range of 0–1 kHz by the overall energy in the frequency range of 0–4 kHz. This measure of spectral tilt is considered as an indicator of emphasis, since in weaker speech segments, energy is more concentrated in the lower frequencies (Tavi & Werner, 2020).

2.3 Functional data analyses

In the first step of FDA, scalar SPI values were transformed into logarithmic continuous functions, or functional SPIs (fSPIs). A (natural) logarithmic scale was used due to the fact that speech perception is also logarithmic (Reetz, 2009). Only the SPI measurements were used in functional analyses because they present all the main prosodic features (Tavi & Werner, 2020) in a single measure. The B-spline basis system was used for transforming the scalar SPI values to fSPIs, as it is a common choice for aperiodic signals (Gubian et al., 2015). The order of polynomial segments was set to four and the number of basis functions was set to 42. The lambda parameter, which defines the amount of smoothness, was 0.1. These parameters were chosen based on visual inspections of the resulting functions using different levels of smoothing. A strong smoothing was applied in order to take into account only the major variation in prosodic prominence and to exclude short-term fluctuation in adjacent syllabic units.

Because the aim of this study was to analyse prosodic prominence in different speaking conditions rather than specific linguistic phrases, the mean fSPI of the four fables was calculated for each speaker for both regular and fast speech. As a result, the mean fSPIs carry information on averaged wide-scale prosodic events, in which the amount of intra-speaker and linguistic variation has been reduced.

Finally, fPCA was applied on the mean fSPIs (hereafter referred to as fSPIs instead of mean fSPIs). Using fPCA, fSPIs can be reconstructed with the formula:

$$f(t) \approx \mu(t) + \sum_{i=1}^{n} s_i \times PC_i(t),$$

where $\mu(t)$ is the mean of the fSPIs, PC_i is the ith principal component function, and s_i is its weight, or score. Because the individual scores model the shape of each function, they can be used to investigate the dynamics of continuous speech features, such as f0 or formant curves (Gubian et al., 2015).

3. Results

3.1 Prosodic features

A total of five prosodic features were extracted from the syllabic units. The measurements were compared using paired T-tests. Articulation rate was measured in order to confirm that syllabic units are truly spoken faster in fast compared to regular speech.

Table 1. Prosodic differences between regular and fast speaking conditions presented as mean values and p-values from paired T-tests (For 'feature' see Section 2.2).

feature	*female speech*					*male speech*				
	regular	*fast*	*t*	*df*	*p*	*regular*	*fast*	*t*	*df*	*p*
AR	5.06	5.86	11.35	15	<.001	4.95	5.95	−14.36	19	<.001
f0 (Hz)	199	190	4.27	15	<.001	120	110	6.17	19	<.001
Eb1kHz	0.91	0.99	−8.74	15	<.001	0.93	0.99	−10.86	19	<.001
dur (s)	0.20	0.17	11.46	15	<.001	0.20	0.17	15.63	19	<.001
SPI	9.19	7.77	8.50	15	<.001	5.51	4.45	13.30	19	<.001

Table 1 shows the mean values of the two speaking conditions calculated from all speakers and the results from the paired T-tests. The results are presented separately for male and female speakers. In comparison to regular speaking condition, speaking fast increased the articulation rate and energy proportion below 1 kHz. Duration, f0 and SPI in syllabic units were decreased. The prosodic differences between the speaking conditions were statistically significant for both the male and the female speakers using a Benjamini & Hochberg -adjusted significance level of 0.05. The results confirmed that syllabic units were spoken faster in fast speech and that increasing speech tempo has a significantly decreasing effect on prosodic prominence.

3.2 Functional syllabic prosody index

In order to examine the dynamic changes in prosodic prominence caused by fast speaking, the SPI trajectories were converted to functional SPIs (see Section 2.3). Figure 1 shows the mean fSPIs for male and female speakers. The clearest difference between the regular and the fast mean fSPIs for both male and female speakers is that the regular fSPIs are above the fast fSPIs. The lower fSPIs for fast speech were expected based on the results of the acoustic-phonetic analyses presented in Section 3.1. The positions of the mean fSPIs are rather consistent throughout the whole functions, indicating that speakers are able to retain constant prominence levels in long speech segments of different tempi.

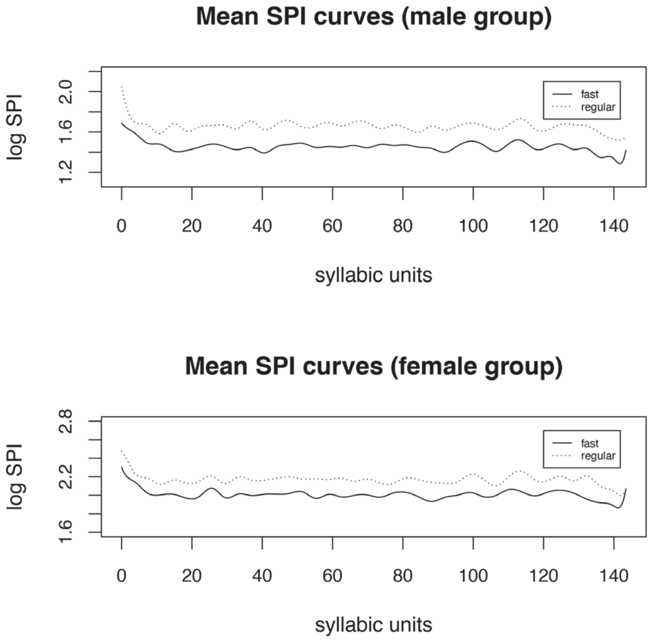

Figure 1. Mean fSPIs for male and female speakers.

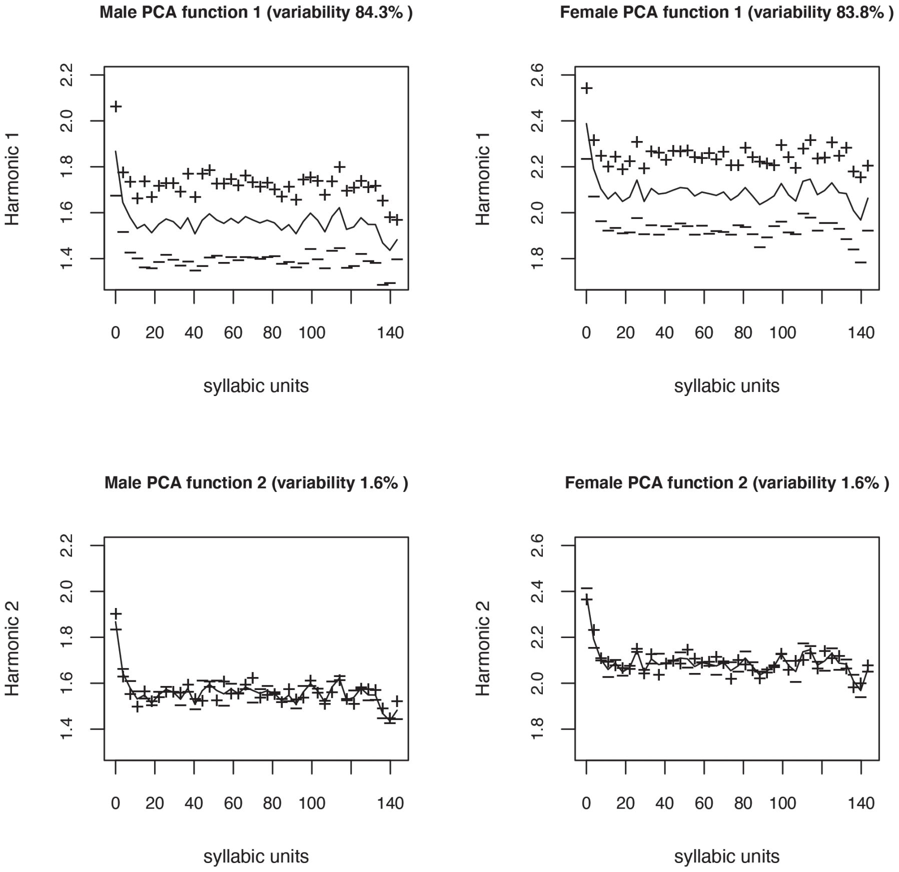

Figure 2. The effect of the first two PC functions on the male and the female speakers' mean fSPIs. Trajectories of plus and minus signs demonstrate the effects of the PC functions and standard deviation of their scores on the mean fSPIs (solid lines).

fPCA was applied to the fSPIs in order to examine the major modes of prominence variation. Figure 2 demonstrates the effects of PC functions on the mean fSPIs using the trajectories of plus and minus signs. These trajectories are formed by multiplying the PC functions by the standard deviation of their weightings, which are then either added to or subtracted from the mean fSPI. The effects are rather similar for male and female speakers: PC1 (top panels) explains the variation related to location of the fSPI. It also explains a major part of the fSPI variability (>80%). An increase of PC1 weighting, or s_1, will raise the mean fSPI, whereas decrease of the weighting will lower it. PC2 and its weighting, or s_2, are more associated with the timings of positive and negative prominence peaks; however, the second PCA function explains only 1.6% of the fSPI variation.

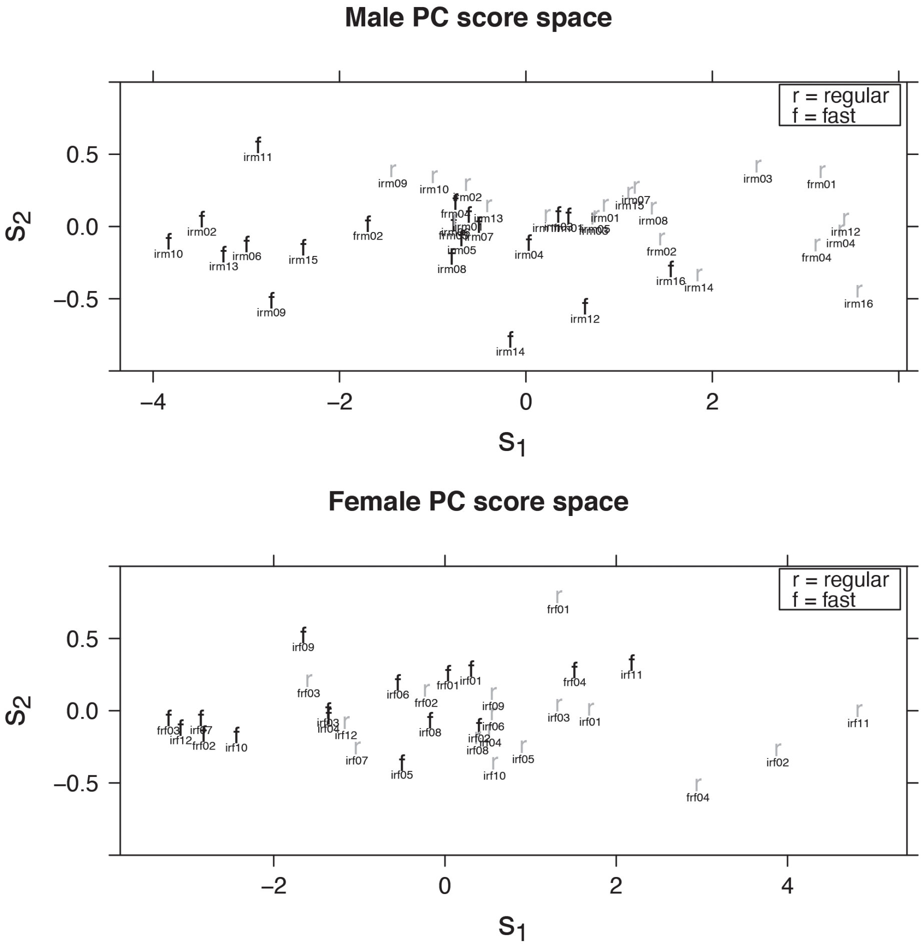

Figure 3. PC score spaces for male and female speakers.

Figure 3 reveals how the two speaking conditions of individual speakers are located in the PC1–PC2 score space. Letters *f* (fast) and *r* (regular) indicate the speaking condition and the identifier below the letters indicates speaker identity. In the PC score spaces, fast speech is mainly located on the left and regular speech on the right. The division between the speaking conditions is clearer with the male speakers, as the female speakers' scores have more overlap. However, each speaker's s_1 of fast speech is lower compared to their s_1 of regular speech (see Figure 4). The s_2 shows no relationship specific to the two speaking conditions.

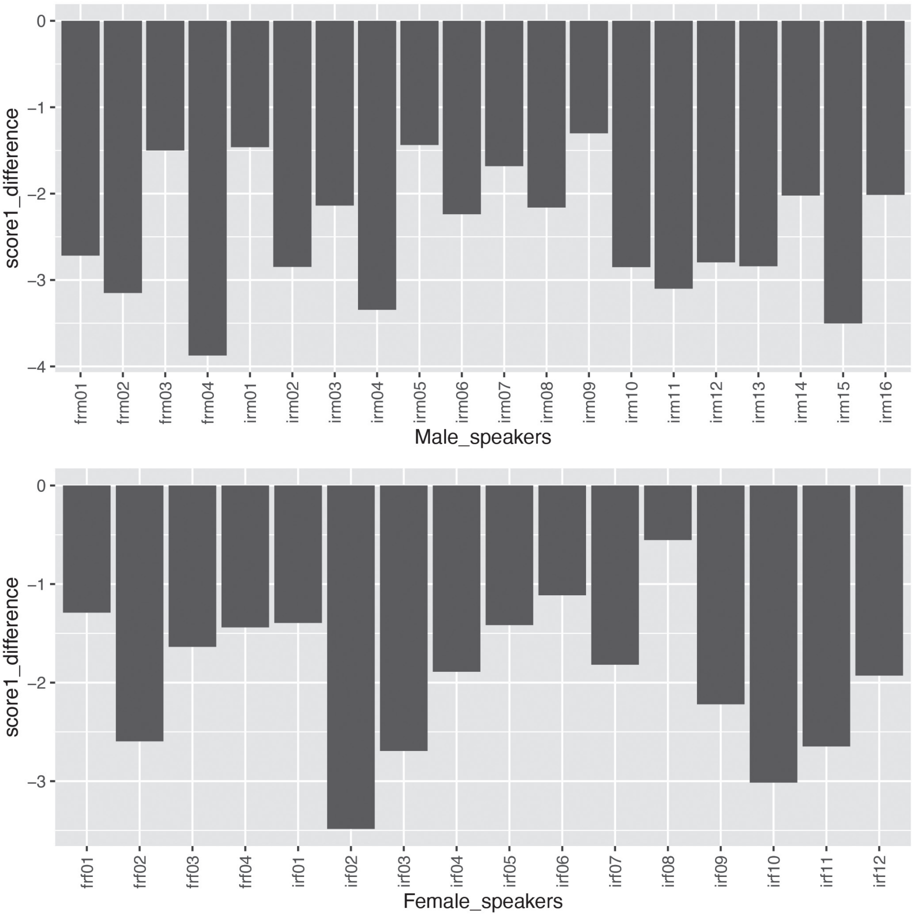

Figure 4. The s_1 differences between fast and regular speech for each speaker. The s_1 of regular speech is subtracted from the s_1 of fast speech.

3.3 Speech intelligibility

Sections 3.1 and 3.2 described several changes in prosody caused by the increase of speech tempo. To test whether these changes are related to speech intelligibility, the four fable passages were transcribed using an ASR system provided by BAS Web Services (Kisler et al., 2017). Then, a Python script[1] was used to calculate WERs for each speaker's regular and fast versions of the passages. WER divides the number of errors (i.e., the substitutions, insertions and deletions) by the total number of words. Although WER is reported as a percentage, it can be more than 100%, because the number of errors can be higher than the number of words in the reference text.

[1] https://holianh.github.io/portfolio/Cach-tinh-WER/

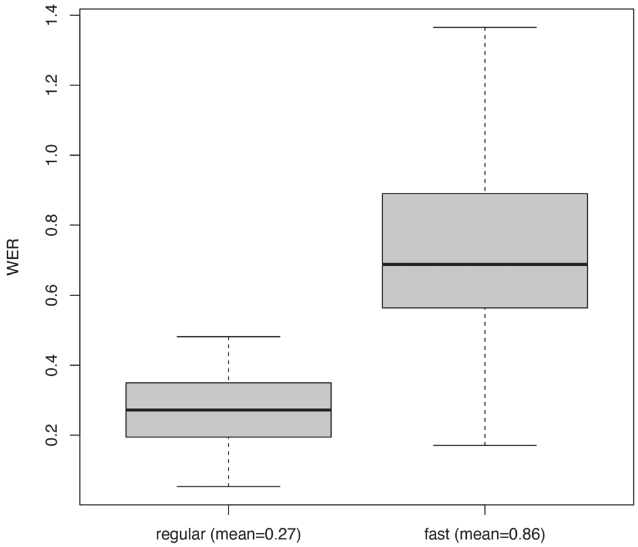

Figure 5. WERs in the regular and the fast-speaking conditions. Two outliers of the fast speech group (WERs = 4.21 and 2.32) were excluded from the figure.

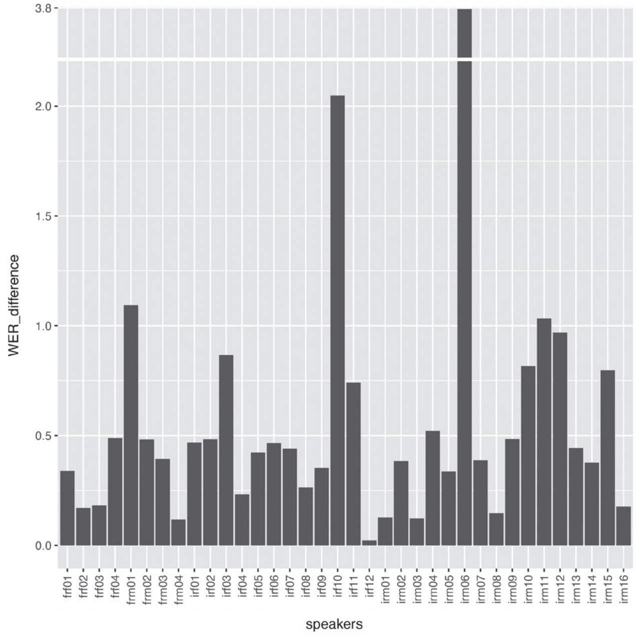

Figure 6. WER differences between fast and regular speech for each speaker. The differences are calculated by subtracting speaker-specific WERs of regular speech from those of fast speech.

Figure 5 reveals a drastic increase of WER in fast speech; whereas the mean WER is 0.27 in regular speech, in fast speech it increases to 0.86. The differences between the two speaking conditions for individual speakers are presented in Figure 6. This result shows that fast speech has a negative effect on ASR accuracy for all speakers, although the amount of difference varies between speakers.

To evaluate the relationship between changes in the prosodic features and speech intelligibility, six correlation tests were carried out using a Benjamini & Hochberg -adjusted significance level of 0.05. The results are presented in Table 2.

Table 2. Correlations between mean values of WER and the mean prosodic features of the speakers. The correlation tests included both speaking conditions. Statistically significant correlations are marked using bold type.

feature	*male speakers*		*female speakers*	
	r	*p*	*r*	*p*
s_1	**–0.48**	**0.004**	–0.31	0.088
SPI	**–0.46**	**0.006**	–0.30	0.102
f0 (Hz)	–0.29	0.088	–0.06	0.725
eb1kHz	**0.43**	**0.008**	**0.49**	**0.008**
duration (s)	**–0.57**	**0.001**	**–0.60**	**0.001**
AR	**0.62**	**>0.001**	**0.64**	**0.001**

The correlation between the s_1 and WERs was statistically significant for the male speakers (r=–0.48), but not for the female speakers (r=–0.31). Similarly, the correlation between the scalar mean SPIs and WERs was statistically significant for the male (r= –0.46) but not for the female speakers (r=–0.30). Therefore, the SPI-related correlations demonstrate at least partial association between decreased prominence and speech intelligibility.

The energy values, syllable durations and ARs also correlated with WERs, which shows that lower speech intelligibility is associated with higher energy proportion below 1 kHz at a faster speech tempo. The correlation was especially strong between AR and WERs (r=0.62 and 0.64), demonstrating the strong negative effect of speaking fast on speech intelligibility. However, there was no significant correlation between f0 and WERs. Overall, the results considering the relationship between prosodic features and WERs were largely similar for the male and the female speakers.

4. Discussion

In the Introduction, two hypotheses were presented: speaking fast (1) decreases prosodic prominence and (2) deteriorates speech intelligibility. The results confirmed both of them. The mean values of articulation rate, syllabic duration, f0, energy proportion below 1 kHz and SPI revealed a significant change towards lower prominence when the speakers spoke fast compared to regular speech tempo.

The dynamic changes in prosodic prominence were investigated using fPCAs, which revealed the major modes of variation in the fSPIs. The changes between the two speaking conditions were first examined for the male and the female speaker groups and then for the individual speakers in the PC score spaces. The first PC and the s_1, which explained over 80% of prominence variation, were mainly related to the overall height of the mean fSPIs. The second PC and the s_2, which explained only 1.6% of the variation, were more associated with the timings of the peaks in the mean fSPIs. Even though the clusters of the two speaking conditions partly overlapped in the PC score spaces, each speaker's s_1 was systematically lower in fast speech, indicating lower fSPIs. The s_2 variation was found to be unrelated to the speaking conditions. Thus, the functional results mainly supported the findings from the conventional prosodic analyses, but also showed rather high inter-speaker variation in prosodic prominence. Moreover, they verified that prosodic prominence is consistently (dynamically) lower in fast speech, which would not have been possible using conventional statistics.

Finally, the negative effects of speaking fast on speech intelligibility were established; in terms of mean WERs, the ASR accuracy decreased drastically from 0.27 to 0.86 when the speakers spoke fast. Even though the amount of decrease in WER varied between the speakers, ASR accuracy decreased for every speaker during fast speech. In addition, there was a statistically significant correlation between the WERs and most of the studied prosodic features.

Overall, this study has shown that when speakers intentionally alter their articulation rate, this has a holistic effect on speech prosody. Hence, the results suggest that it might be difficult, or even impossible, for speakers to alter speech tempo without an impact on other prosodic features. One of the few exceptions according to the previous literature might be speech rhythm, which was shown to have no significant within-speaker variation in different tempo conditions (Dellwo et al., 2015). Nevertheless, if the implication above holds, different prosodic aspects of speech can be even more connected than has been assumed in previous studies. Therefore, an aim for future studies would be to verify whether or not speakers are capable of conducting only tempo-related changes in speech prosody. In order to achieve this aim, functional data analyses can provide an efficient methodological framework.

5. Conclusions

In this study, prosodic changes caused by an increase of speech tempo were investigated. Dynamic changes in prosodic prominence were studied using SPI, a novel prominence measure, and functional PCA. In addition, the effects of increased tempo on speech intelligibility were evaluated using ASR. The results confirmed an expected increase in articulation rate and decrease in syllable duration in fast speech. In addition, energy proportion below 1 kHz was found to increase and f0 and SPI to decrease. fPCA verified dynamic changes in the functional SPIs, showing a systematic decrease for each speaker in fast speech. Finally, automatic transcriptions using ASR substantiated the negative effect of speaking fast on speech intelligibility. In addition, most of the prosodic measures correlated with the ASR accuracy.

REFERENCES

Boersma, P. & Weenink, D. (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.32. url: http://www.praat.org.

Corretge, R. (2020). Praat Vocal Toolkit. url: http://www.praatvocaltoolkit.com.

Cronenberg, J., Gubian, M., Harrington, J., & Ruch, H. (2020). A dynamic model of the change from pre-to post-aspiration in Andalusian Spanish. *Journal of Phonetics*, 83, 1–22. doi: 10.1016/j.wocn.2020.101016.

Cummins, F., Grimaldi, M., Leonard, T., Simko, J. (2006). The chains corpus: Characterizing individual speakers. *Proceedings of SPECOM*, Citeseer, pp. 431–435.

De Jong, N. H. & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behaviour Research Methods, 41*, 385–390.

Dellwo, V., Leemann, A., & Kolly, M. J. (2015). Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *The Journal of the Acoustical Society of America*, 137(3), 1513–1528.

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech – a syllable-centric perspective. *Journal of Phonetics*, 31, 465–485.

Gubian, M., Boves, L., & Cangemi, F. (2011). Joint analysis of f0 and speech rate with functional data analysis. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* IEEE, Prague, Czech Republic, pp. 4972–4975.

Gubian, M., Cangemi, F., & Boves, L. (2010). Automatic and data driven pitch contour manipulation with functional data analysis. *Speech Prosody*, Chicago, IL, USA.

Gubian, M., Torreira, F., & Boves, L. (2015). Using functional data analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics*, 49, 16–40. doi:10.1016/j.wocn.2014.10.001.

Hazan, V. & Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America*, 116, 3108–3118.

Janse, E. (2004). Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech. *Speech Communication*, 42, 155–173. doi: https://doi.org/10.1016/j.specom.2003.07.001.

Janse, E., Nooteboom, S., & Quené, H. (2003). Word-level intelligibility of time-compressed speech: prosodic and segmental factors. *Speech Communication*, 41, 287–301.

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language,* 45, 326–347. doi: http://dx.doi.org/10.1016/j.csl.2017.01.005.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the h&h theory. *Speech production and speech modelling*, Springer, pp. 403–439.

Mayo, C., Aubanel, V., & Cooke, M. (2012). Effect of prosodic changes on speech intelligibility. *Proceedings of INTERSPEECH*, Portland, USA September 9-13, pp. 1708–1711.

Niebuhr, O., & Kohler, K. J. (2011). Perception of phonetic detail in the identification of highly reduced words. *Journal of Phonetics*, 39(3), 319–329.

Patel, R. & Schell, K.W. (2008). The influence of linguistic content on the lombard effect. *Journal of Speech, Language, and Hearing Research*, 51(1), 209–220. doi: 10.1044/1092-4388(2008/016)

Ramsay, J.O., Hooker, G., & Graves, S. (2009). *Functional data analysis with R and MATLAB*. NY: Springer.

Reetz, H. (2009). *Phonetics: transcription, production, acoustics, and perception.* Oxford: Wiley-Blackwell.

Roettger, T.B., Winter, B., & Baayen, H. (2019). Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Journal of Phonetics,* 73, 1–7.

Stanton, B., Jamieson, L. & Allen, G. (1988). Acoustic-phonetic analysis of loud and lombard speech in simulated cockpit conditions. *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pp. 331–334. doi: 10.1109/ICASSP.1988.196583.

Tavi, L. & Werner, S. (2020). A phonetic case study on prosodic variability in suicidal emergency calls. *International Journal of Speech, Language & the Law*, 27, 59–74.

Tomashenko, N., Wang, X., Vincent, E., Patino, J., Srivastava, B. M. L., Noé, P.G., Nautsch, A., Evans, N., Yamagishi, J., O'Brien, B., & Chanclu, A. (2021). The voiceprivacy 2020 challenge: Results and findings. arXiv preprint arXiv:2109.00648.

Zellers, M., Gubian, M., & Post, B. (2010). Redescribing intonational categories with functional data analysis. *Proceedings of INTERSPEECH*, Makuhari, Japan, pp. 1141–1144.

Lauri Tavi
School of Humanities
University of Eastern Finland
Agora, Finland
E-mail: lauri.tavi@uef.fi