INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Schwa Realization in French: Using Automatic Speech Processing to Study Phonological and Socio-linguistic Factors in Large Corpora

Yaru Wu[1], *Martine Adda-Decker*[1,2], *Cécile Fougeron*[1], *Lori Lamel*[2]

[1]Laboratoire de Phonétique et Phonologie (UMR7018, CNRS-Sorbonne Nouvelle), France
[2]LIMSI, CNRS, Université Paris-Saclay, Rue John von Neumann, F-91405 Orsay Cedex, France
yaru.wu@univ-paris3.fr, madda@limsi.fr, cecile.fougeron@univ-paris3.fr, lamel@limsi.fr

Abstract

The study investigates different factors influencing schwa realization in French: phonological factors, speech style, gender, and socio-professional status. Three large corpora, two of public journalistic speech (ESTER and ETAPE) and one of casual speech (NCCFr) are used. The absence/presence of schwa is automatically decided via forced alignment, which has a successful performance rate of 95%. Only polysyllabic words including a potential schwa in the word-initial syllable are studied in order to control for variability in word structure and position. The effect of the left context, grouped into classes of a word final vowel or final consonant or a pause, is studied. Words preceded by a vowel (V#) tend to favor schwa deletion. Interestingly, words preceded by a consonant or a pause have similar behaviors: speakers tend to maintain schwa in both contexts. As can be expected, the more casual the speech, the more frequently schwa is dropped. Males tend to delete more schwas than females, and journalists are more likely to delete schwa than politicians. These results suggest that beyond phonology, other factors such as gender, style and socio-professional status influence the realization of schwa.

Index Terms: schwa, large corpora, forced alignment, speech style, pre-boundary context, socio-linguistic factor, mixed model

1. Introduction

The realization/deletion of schwa, a vowel that alternates with ∅, is considered to be one of the most complicated phenomena in French phonology. Since Grammont [1], schwa is described as an unstable vowel whose behavior depends particularly on its consonantal environment. According to Grammont, the realization of schwa becomes an obligation for most of the cases when the surface form resulting from the deletion of schwa would have three or more consonants in a row. Grammont thus proposed the Three Consonants Rule ('Loi des 3 consonnes'). The Three Consonants Rule has since then been explored and restudied by numerous phoneticians and phonologists (see for e.g. Durand [2]). However, the phenomenon is not that simple in real-life production.

Research on schwa realization in very large corpora was beyond imagination before the availability of automatic speech processing to help provide initial annotations and their temporal position in the audio. Our aim is to investigate both phonological and socio-linguistic factors on schwa realization. Prosody is considered to potentially have an impact on schwa realization [3, 4]. The impact of prosody is not discussed in this work since there is no prosodic information annotated in our data. Our goal is to propose a method for linguistic analyses using tools from automatic speech processing and to evaluate the importance of socio-linguistic investigations on phonological phenomena.

2. Corpora and alignment

Three large corpora containing spontaneous speech are used: ESTER [5], ETAPE [6] and NCCFr [7]. The ESTER corpus contains 100 hours of radio broadcast news shows in French. The ETAPE corpus consists of 13.5 hours of radio data and 29 hours of TV data in French, including debates and free conversations. The Nijmegen Corpus of Casual French (NCCFr) is composed of 35 hours of casual French conversation between friends.

The data are automatically segmented into words and phonemes using the LIMSI speech transcription system [8] in forced alignment mode. Forced alignment uses an acoustic model to find the optimal association of speech segments with a phonemic transcription (obtained via a pronunciation lexicon) corresponding to the word level transcription of the segment. The orthographic transcriptions is provided to the LIMSI transcription system which is used in alignment mode and returns the word and phone boundaries. The system automatically chooses the most adapted pronunciation for each word from the possible pronunciations in the dictionary. Pause, hesitation or breath are also detected automatically by the system. The minimum duration of a segment is 30ms, corresponding to 3 frames [9].

The combination of forced alignment and pronunciation dictionaries including specific variants may serve as a tool to study possible emerging tendencies of speech and analyse big speech data [10]. In order to quantify the variants in question, forced alignment is used to locate words and to select their best variants. Using this approach, schwa realization can be automatically decided via forced alignment as soon as the corresponding variants concerning schwa are present in the pronunciation dictionary. The used approach is therefore very different from the one of the PFC project [11], for instance, where the presence and absence of schwa is determined through auditory perception.

As the alignment system may chose the best matching variants, the resulting segmental transcriptions tend to be close to the pronunciation of the speakers. Figure 1 demonstrates the spectrograms of the word "dessus"(/dəsy/, *above*) with (Figure 1 (a), [dəsy]) and without (Figure 1 (b), [dsy]) schwa aligned. To verify the accuracy of the automatic alignment, a subset of the segments were manually checked for the absence or presence of schwa. The successful performance rate of the automatic alignment is 95%.

Methods and tools like forced alignment initially developed for automatic speech recognition may be very helpful for investigating linguistic hypotheses in production. The method we

http://dx.doi.org/10.21437/Interspeech.2017-470

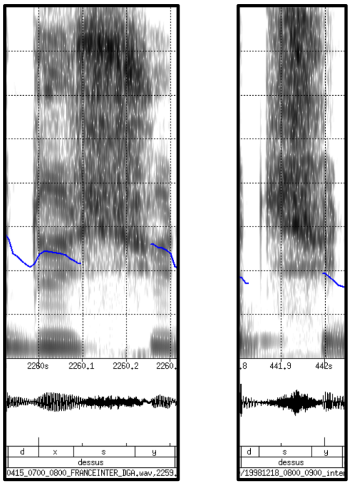

Figure 1: *The word "dessus" (/dəsy/, above) with (a) and without (b) schwa aligned by the LIMSI speech transcription system.*

developed using automatic speech recognition can help us test different linguistic hypothesis.

3. Method

In order to better control for word structure and position related variability, only polysyllabic (2 or more syllables) words of type CəCV (e.g. semaine /səmɛn/: *week*) and of type CəCCV (e.g.retraité /rətrɛte/: *retired*), i.e. including a potential schwa in word-initial position, are chosen for our analyses. Note that word of type CCəCV (e.g. prenons /prənɔ̃/: *second person plural form for "take"*) had to be discarded since they were restricted to a limited set of words.

Speech segments corresponding to the words of the type CəCV and the type CəCVV are selected, and the corresponding information (i.e. aligned pronunciation, left context of the word in question, corpus name, name of the TV show with recording date, speaker, gender and socio-professional status) is extracted. We discarded speakers who had fewer than 50 occurrences of CəCV and CəCCV word-tokens (leaving 158 speakers to be included in this study). In order to determine the absence or presence of schwa, we decided to use the phonological transcription of Lexique380 [12] as a reference. The transcription is therefore used as a pronunciation reference for our analyses. That is to say, the presence or absence of schwa is determined by comparing the transcription of Lexique (canonical form) and the aligned forms of the three corpora ESTER, ETAPE, NCCFr (production of speakers). It should be pointed out that every polysyllabic word with potential schwa in word initial syllable (i.e. every word we analyze) has a reference in Lexique380. Thus, each word in ESTER, ETAPE or NCCFr that we analyze has both an aligned pronunciation and a canonical pronunciation. This additional information concerning the canonical pronunciation allows us to categorize words in question into "schwa present" and "schwa absent". The categories are used for our statistical analyses.

The canonical pronunciations of the words in question are also “mapped" according to the nature of the consonants. The consonants are classified into eight categories: voiceless stop (**o**), voiced stop (**O**), voiceless fricative (**f**), voiced fricative (**F**), nasal (**N**), liquid (**L**), glide (**G**) and full vowel (**V**). For instance, the word “semaine" (/səmɛn/, *week*) is assigned a phonological annotation “**fəNVN**". This additional information is used in our statistical analyses in Section 5.

Table 1: *Overview of the factors in question. "#" stands for word boundary ; "_" stands for the position of the word in question; "V" stands for full vowel (schwa excluded); "O#" stands for pause. Pause refers to segments that are silence, hesitations and breaths. “␣" refers to segments that are O#.*

Category	factor	detail
phono-logical	word form ("type")	cəcv (e.g.semaine /s(ə)mɛn/: *week*)
		cəccv (e.g. retraité /r(ə)trɛte/: *retired*)
	pre-boundary context	V#_ (e.g.la#semaine /la#s(ə)mɛn/: *the week*)
		C#_ (e.g.cette#semaine /sɛt#s(ə)mɛn/: *this week*)
		O#_ (e.g.␣Regarde! /␣r(ə)gard/: *Look!*)
socio-linguistic	speech style (“corpus")	ESTER (formal journalistic speech)
		ETAPE (casual journalistic speech)
		NCCFr (conversational speech)
	gender	male
		female
	career	journalist
		politician
other-factors mentioned	speaker	/
	consonant nature	e.g. semaine /s(ə)mɛn/: *week* ==>**fəNVN** (voiceless fricative + ə + nasal + full vowel + nasal)
	recording session	e.g. BFMTV_BFMStory_2010_09_14_175900 => the TV show BFMStory of BFMTV, recorded on Sep. 14th 2010

4. Phonological and sociolinguistic factors

This section presents the phonological and socio-lingustic factors chosen for this study. Table 1 summarizes the conditions analysed and the notation used in this article.

Concerning the phonological factors, we analyse word forms and the left context of the words. Word forms (coded as "type" for the statistical analyses and the figures) indicate words of type CəCV and of type CəCCV that we chose for this study. We expect CəCV words to favor schwa deletion more than CəCCV words (according to the Three Consonants Rule of Grammont). We are particularly interested in the left contexts of the words in question (pre-boundary contexts). The pre-boundary contexts (coded as "context") are:

- V#_
- C#_
- O#_

with “#" as word boundary, “_" as position of the word in question, “V" as full vowel (which does not include schwa), “C" as consonant and “O" as pause. Pause includes all segments that are silent pauses, hesitations and breaths. Words preceded by a vowel are expected to facilitate schwa deletion more than those preceded by a consonant, restrained by the Three Consonants Rule. Côté [13] and many others observe that schwa after a post-pausal position (e.g. demande-la /dəmɑ̃d la/: *request it*; je suis /ʒə sɥi/: *I am*) is optional. We are interested in finding out how schwa would behave in words preceded by O# in our data.

As mentioned earlier, speech styles, gender and socio-profesionnal status are chosen as our socio-linguistic factors. As regards the speech styles (coded as "corpus"), we suppose that schwa deletion occurs more frequently in NCCFr than in ETAPE and ESTER, since NCCFr is consist of conversational spontaneous speech. As for ESTER and ETAPE, schwa is expected to be omitted more in ETAPE than in ESTER, given that ETAPE contains more data of free conversations and debates on television.

Gender influence (also coded as "gender") is also analyzed. As is well known that male speakers articulate with less precision than female speakers (see for e.g. [14]), we want to verify whether male speakers deletes schwa more than female speakers.

Analyses on the effect of social-professional status (coded as "career") are limited to the ETAPE corpus to better control for speech style related variation. Since most of our speakers in the ETAPE corpus are either journalists or politicians, we are interested in finding out whether journalists and politicians have different behaviors on schwa realization.We suppose that journalists tend to maintain less schwa then politicians since journalists are more aware of limited time slots they are allocated when they speak in public.

5. Analyses and Results

Generalized linear mixed models are used for the statistical analyses of this study [15]. The effects of both phonological and socio-linguistic factors are analysed using the package lme4 [16] in R [17].

The fixed factors considered were: corpus (ETAPE, ESTER or NCCFr, reference: ESTER), type (CəCV vs. CəCCV, reference: CəCCV), context (O#, C# or V# , reference: O#) and gender (male or female, reference: female). The following random terms were included in the model: a random intercept per speaker and one per consonant nature (which was mentioned in Section 3); by-speaker slopes for the effect of type and context; by-consonant-nature slopes for the effect of context and gender. Details on the factors can be found in Table 1. Post-hoc tests based on the model were performed for fixed effects to get information on each level of each fixed effect.

The probability to observe a schwa decreases significantly both in ETAPE (log odds ratio = -0.546, |Z|=4.416, p<0.001), and in NCCFr (log odds ratio = -2.798, |Z|=20.499, p<0.001) with respect to that observed in ESTER. Schwa is realized less often when the type is CəCV than when the type is CəCCV (log odds ratio = -0.719, |Z|=3.528, p<0.001). Both C# and V# contexts have a significantly negative effect on the realization of schwa with respect to that observed in O# (C#: log odds ratio = -0.834, |Z|=2.560, p<0.05; V#: log odds ratio = -1.887, |Z|=5.905, p<0.001). Finally, male speakers realize less schwa than female speakers (log odds ratio = -0.339, |Z|=2.445, p<0.05).

Figures 2 to 5 illustrate the probability of the presence of schwa base on the model above. A different scale is needed for Figure 4 in order to show the effect of the corpus on the probability that schwa is produced. Note that the overlap of the error bars we see on these figures doesn't necessarily indicate the non-significance between the variables, on account of the fact that the standard error of a model coefficient differs from that of a model's predicted value. Details can be found in [18].

Figure 2 shows the probability of the presence of schwa as a function of "type" (i.e. words of type CəCCV vs. words of type CəCV). As we can see in Figure 2, words of type CəCCV tend to maintain schwa more than words of type CəCV. The results of our model and that of the post-hoc test deriving from the model indicate that the difference between the two types are significant ($p < 0.001$). The results concerning the word type effect show that the realization of schwa is under the influence of Grammont's Three Consonants Rule. If the deletion of schwa generates a succession of three consonants, schwa is prone to be pronounced.

The probability of the presence of schwa concerning the

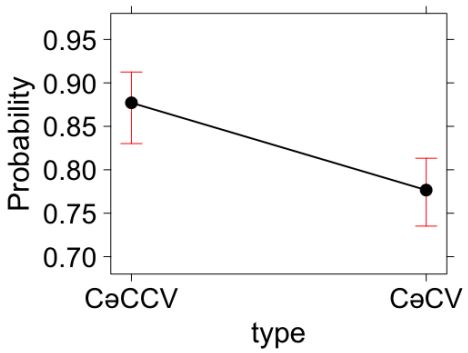

Figure 2: *Probability of the presence of schwa for CəCV and CəCCV word types.*

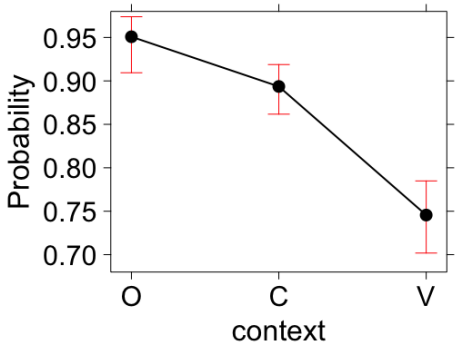

Figure 3: *Probability of the presence of schwa as a function of the pre-boundary context, namely word final vowel (V#) or final consonant (V#) or a pause (O#). Pause refers to segments that are silence, hesitations and breaths. "#" stands for word boundary ; "_" stands for the position of the word in question; "V" stands for full vowel (schwa excluded).*

pre-boundary context (left context of the word) is shown in Figure 3. Words containing schwa preceded by a V# (vowel) tend to facilitate schwa deletion. As expected, words preceded by a consonant tend to maintain schwa since the more consonants the surface form has, the less schwa is omitted (Grammont's Three Consonants Rule). Interestingly, pre-boundary contexts C# and O# tend to impact schwa realization similarly: schwa tends to be present for pre-boundary contexts O# as well. Post-hoc test argues that the difference between the context V# and C# is significant (p<0.001). Likewise, the context V# and the context O# are significantly different (p<0.001). Surprisingly, the context O# has the highest probability for the presence of schwa. As mentioned in Section 4, schwa is considered optional in the literature in the context O#. Our results concerning O# contexts suggest that, monosyllabic words aside, speakers are most likely to maintain word-initial schwa when the pre-boundary context is O# in journalistic and conversational speech for polysyllabic words.

According to Figure 4, it is less likely to have schwa deletion in the corpus ESTER and the corpus ETAPE than in the corpus NCCFr. The post-hoc test based on our model suggests significant differences between the three corpora (ESTER vs. ETAPE: p<0.001; ETAPE vs. NCCFr: p<0.001; ESTER vs. NCCFr: p<0.001). The results concerning the effect of corpus show that speakers are prone to produce less schwa when they are in a less formal situation. Schwa is more likely to be realized in journalistic speech (ESTER or ETAPE) than in conversational speech (NCCFr), and there is a higher possibility for

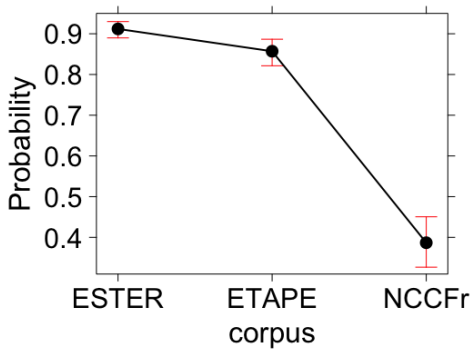

Figure 4: *Probability of the presence of schwa as a function of corpus: ESTER, ETAPE and NCCFr.*

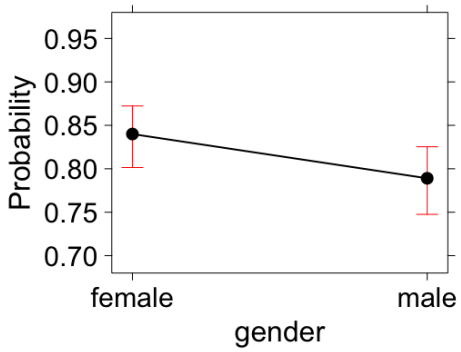

Figure 5: *Probability of the presence of schwa for male and female speakers.*

schwa to be realized for formal journalistic speech (ESTER) than that for casual journalistic speech (ETAPE).

Gender differences are also found in our data according to the model-based post-hoc test ($p<0.05$). Figure 5 shows that male speakers favors omitting schwa more than female speakers. This observation of gender difference is consistent with the observation in [19] where a study of pronunciations in 4000 hours of manually transcribed speech in French and English showed that male speakers use more reduced pronunciations than female speakers and also have a larger proportion of filled pauses and repetitions.

As mentioned earlier, we decided to investigate the effect of career by using the ETAPE corpus only, and we investigate data concerning "journalists" and "politicians" exclusively. Therefore, we performed a separate model to investigate career influence on schwa realization. The fixed factor considered was career (journalist vs. politician, reference: journalist). The following random terms were included in the model: a random intercept per speaker and one per consonant nature; by-speaker slopes for the effect of career; by-consonant-nature slopes for the effect of career. Details on the factors can be found in Table 1. The probability to observe a schwa increases significantly for politicians (log odds ratio = 0.625, $|Z|=2.284$, $p<0.05$) with respect to that observed for journalists. Results concerning the career effect indicate that public speakers take different strategies in speech production, as far as schwa realization is concerned. Politicians tend to produce schwa more often than journalists.

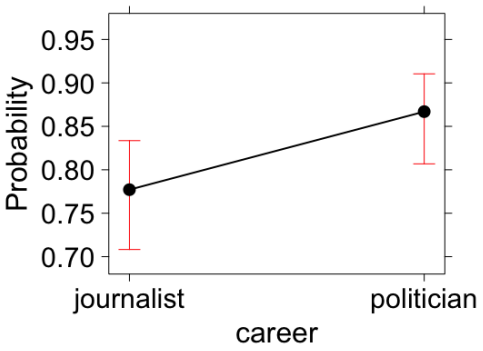

Figure 6: *Probability of the presence of schwa for journalist and politician.*

6. Discussion and Conclusions

Our results on phonological factors are consistent with Grammont's Three Consonants Rule and extend the research to pre-boundary level. We looked at the influence of number of consonants inside a polysyllabic word and we checked different pre-boundary contexts on the left side of the words (i.e. this context is a word final vowel or final consonant or a pause beyond word boundary). Restrained by the Three Consonants Rule, words of type CəCV (2 consonants in a row in the surface form if schwa is deleted) are more likely to drop schwa than words of type CəCCV (3 consonants in a row in the surface form if schwa is deleted). Words preceded by a vowel (V#) tend to favor schwa deletion. By contrast, words preceded by a consonant (C#CəCV or C#CəCCV, 3 or 4 consonants in a row in the surface form if schwa is deleted) or a pause prevent the deletion of schwa. Results concerning the phonological factors show that schwa realization does not only concern the word itself, but also its preceding context. Socio-linguistic factors are seen to have a substantial impact on schwa realization. Our results on speech style reveal that schwa drops more in less formal settings. Schwa is more likely to appear in journalistic speech than in conversational speech. More precisely, formal journalistic speech (ESTER) has a higher chance for schwa to be realized than casual journalistic speech (ETAPE); conversational speech (NCCFr) has a much lower probability for schwa to be realized than both formal journalistic speech and casual journalistic speech. Male speakers tend to delete schwa more than female speakers. Public speakers have different speaking strategies, as far as schwa realization is concerned. Journalists are more likely to delete schwa than politicians. These results suggest that socio-linguistic factors play an important role on schwa realization.

We also develop a research method for phonetic and phonology analyses while using speech technology. The proposed methodology based on forced alignment and pronunciation variants produces results comparable to manual investigations. As it may readily apply to additional data sets, it offers the opportunity to improve and fine-tune the current descriptions on schwa deletion in French. Moreover, the current method is not limited to schwa realization. Other interesting research subjects could also be explored using the same method.

7. Acknowledgements

The present study was funded by the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083).

8. References

[1] M. Grammont, "Le patois de la franche-montagne et en particulier de damprichard (franche-comté). iv: La loi des trois consonnes," *Mémoires de la Société de linguistique de Paris*, vol. 8, pp. 53–90, 1894.

[2] J. Durand, B. Laks, and C. Lyche, "Le projet pfc (phonologie du français contemporain): une source de données primaires structurées," *Phonologie, variation et accents du français. Paris: Hermès*, pp. 19–61, 2009.

[3] M. Swerts and E. Krahmer, "Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions," *Journal of Phonetics*, vol. 38, no. 2, pp. 197–206, 2010.

[4] M. Adda-Decker, R. Nemoto, and J. Durand, "Stratégies de démarcation du mot en français: une étude expérimentale sur grand corpus," *Actes des/Proceedings of JEL*, pp. 91–96, 2009.

[5] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri, "Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news," in *Proceedings of LREC*, vol. 6, 2006, pp. 315–320.

[6] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert, "The etape corpus for the evaluation of speech-based tv content processing in the french language," in *LREC-Eighth international conference on Language Resources and Evaluation*, 2012.

[7] F. Torreira, M. Adda-Decker, and M. Ernestus, "The nijmegen corpus of casual french," *Speech Communication*, vol. 52, no. 3, pp. 201–212, 2010.

[8] J.-L. Gauvain, L. Lamel, and G. Adda, "The limsi broadcast news transcription system," *Speech communication*, vol. 37, no. 1, pp. 89–108, 2002.

[9] M. Adda-Decker and L. Lamel, "The use of lexica in automatic speech recognition," in *Lexicon Development for Speech and Language Processing*. Springer, 2000, pp. 235–266.

[10] M. Adda-Decker, P. Boula de Mareüil, and L. Lamel, "Pronunciation variants in french: schwa & liaison," in *Proceedings of the XIVth International Congress of Phonetic Sciences*, 1999, pp. 2239–2242.

[11] J. Durand, B. Laks, and C. Lyche, "Le projet" phonologie du français contemporain"(pfc)," in *La tribune internationale des langues vivantes*, no. 33, 2003, pp. 3–10.

[12] B. New, M. Brysbaert, J. Veronis, and C. Pallier, "The use of film subtitles to estimate word frequencies," *Applied psycholinguistics*, vol. 28, no. 04, pp. 661–677, 2007.

[13] M.-H. Côté, "Consonant cluster phonotactics: a perceptual approach," Ph.D. dissertation, Massachusetts Institute of Technology, 2000.

[14] A. P. Simpson, "Gender-specific differences in the articulatory and acoustic realization of interword vowel sequences in american english," in *5th Seminar on Speech Production: Models and Data. Kloster Seeon*. Citeseer, 2000, pp. 209–212.

[15] C. E. McCulloch and J. M. Neuhaus, *Generalized linear mixed models*. Wiley Online Library, 2001.

[16] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[17] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016. [Online]. Available: https://www.R-project.org/

[18] J. Fox, "Effect displays in r for generalised linear models," *Journal of statistical software*, vol. 8, no. 15, pp. 1–27, 2003.

[19] M. Adda-Decker and L. Lamel, "Do speech recognizers prefer female speakers?" in *INTERSPEECH*, 2005, pp. 2205–2208.