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

Developing an Embosi (Bantu C25) Speech Variant Dictionary to Model Vowel Elision and Morpheme Deletion

Jamison Cooper-Leavitt[1], *Lori Lamel*[1], *Annie Rialland*[2], *Martine Adda-Decker*[1,2], *Gilles Adda*[1]

[1]LIMSI, CNRS, Université Paris-Saclay, France
[2]LPP, CNRS-Paris 3/Sorbonne Nouvelle, France

cooperleavitt@limsi.fr, lamel@limsi.fr, annie.rialland@univ-paris3.fr,
Martine.Adda@limsi.fr, Gilles.Adda@limsi.fr

Abstract

This paper investigates vowel elision and morpheme deletion in Embosi (Bantu C25), an under-resourced language spoken in the Republic of Congo. We propose that the observed morpheme deletion is morphological, and that vowel elision is phonological. The study focuses on vowel elision that occurs across word boundaries between the contact of long/short vowels (i.e. CV[long] # V[short].CV), and between the contact of short/short vowels (CV[short] # V[short].CV). Several different categories of morphemes are explored: (i) prepositions (*ya*, *mo*), (ii) class-noun nominal prefixes (*ba*, etc.), (iii) singular subject pronouns (*ngá*, *nɔ*, *wa*). For example, the preposition, *ya*, regularly deletes allowing for vowel elision if vowel contact occurs between the head of the noun phrase and the previous word. Phonetically motivated speech variants are proposed in the lexicon used for forced alignment (segmentation) enabling these phenomena to be quantified in the corpus so as to develop a dictionary containing relevant phonetic variants.

Index Terms: phonetics, phonology, language modeling, under-resourced languages

1. Introduction

This paper attempts to quantify certain phenomena in Embosi speech using the forced alignment (segmentation) of an Embosi speech corpus. Embosi is an under-resourced language, and applying a speech tool such as forced alignment can be a valuable approach in investigating phonological effects such as vowel elision and morpheme deletion.

Embosi is a Bantu (C25) language, which is spoken in the Cuvette region and in Brazzaville, Republic of Congo. A speech corpus has been developed for Embosi. The source of the data comes from the Embosi corpus of the Breaking the Unwritten Language Barrier (BULB) project [1]. The observations concerning the language were done via a forced alignment segmentation of phonemes to phones using LIMSI's STK speech processing tool kit [2, 3].

For this paper, vowel elision and morpheme deletion [4] are explored within the context of the Embosi speech corpus and by using the STK speech processing tools. Plenty of examples made from the speech corpus support arguments claiming vowel elision has a strong effect on the phonetic representation of Embosi [5]. Another pattern that has manifested in the speech corpus is morpheme deletion. For example, in the speech corpus the associative preposition *ya*, 'of', is deleted [6]. We examine this morpheme deletion as to how it affects vowel elision, and the methods we use to quantify its representation in the speech corpus. This paper will briefly discuss two topics: (i) the speech corpus used as the primary source of data and examples; (ii) a dictionary of speech variants developed to represent lexical items in the speech corpus. These topics will lead to further investigation in answering the following question: What methods and lexical representations are necessary to model vowel elision and morpheme deletion in Embosi?

2. Methods and Tools

The BULB project aims to document unwritten languages using automatic speech recognition (ASR) and machine translation (MT) tools [7]. The project's focus has been on three unwritten African languages (Embosi, Bassa and Myene). Two phases of the project exist: The first is to collect a large corpus of speech for each language (100 hours) in the form of elicited speech, stories, dialogues, broadcasts, etc. Re-speaking is performed on the collected speech materials, which is followed by an oral translation into French. The next phase of the project is to apply ASR to create phonetic transcriptions of the speech corpora, and to apply MT to create meaningful alignments between the source languages and the target language (French).

2.1. Tools

The corpus was recorded using LIG-AIKUMA (Aikuma) [8], a tablet based speech software application used to make speech recordings during language based field work. The method used in the recording of the audio files was done with the initial utterance of the speaker, followed by a careful re-speaking, and then a French translation of the original utterance. The sound files (careful re-speaking) were transcribed by native Embosi speakers.

2.2. Corpus

This paper makes use of a small speech corpus of Embosi which is currently being developed under the BULB project. The project's objective is to collect large volumes of data from dozens of speakers in different speaking styles. Practically, two 1-month field trips to Brazzaville, Congo have been completed by a native Embosi speaker using a tablet installed with the Aikuma software. So far, within the BULB project, 48 hours of speech data of the Embosi language have been collected in Brazzaville, Congo. The corpus is composed of 50 hours of different styles, from which 4.5 hours have been used for the study described here.

This sub-corpus is composed of elicited speech (read by 3 male Embosi speakers) in two forms: (i) 1472 sentences, extracted from reference sentences for oral language documentation [9] have been translated and written in Embosi (1.3 hours); (ii) 3706 sentences, extracted from an Embosi dictionary [10]. There are a total of 5178 individual utterances where each is saved as a separate audio file in the corpus.

http://dx.doi.org/10.21437/Interspeech.2017-1280

The segmentation model which was done using LIMSI's STK speech analysis software identified a set of 68 separate phones in the corpus. The phone pairs in Table 1 represent 68 Latin-1 phone symbol to IPA symbol pairs organized into three columns. (An additional symbol ' . ' not shown in Table 1 was also used for silence in the segmentation model.) The pairs are shown in descending order of most occurrences to the least occurrences in the speech alignment of the acoustic signal to the Latin-1 phone symbols. There were a total of 96,182 phones aligned to the acoustic signal in the training data. (This total does not include alignments made for silence.) The top left symbol pair, 'Â, á' occurred most frequently and the bottom right symbol pair, 'ú, uú' occurred least frequently. For the ASR to make phonetic transcriptions, phonetic categories must be mapped to a single Latin-1 symbol. For consonants, this was a straightforward one-to-one mapping of an IPA phonetic symbol to a Latin-1 phone symbol. Although, there were certain considerations taken for mapping IPA diacritic symbols to Latin-1 phones. However, mapping vowels proved problematic, since Embosi vowels have not only vowel features such as tongue height and tongue position to consider, but also tone and vowel length. Thus, multiple symbol sets were employed to map vowels of different tone and length as separate categories, as shown in the Table 1.

Table 1: *List of Latin-1 phone symbols and IPA phones used in the ASR model with their total number of occurrences*

Phone symbol to IPA symbol								
Â	á	9607	U	u	1449	f	f	278
A	a	9319	M	^{m}b	1446	µ	^{m}b^{v}	259
l	l	5775	N	^{n}d	1434	P	p^{f}	239
I	i	5471	p	p	1400	i	ii	205
Î	í	5236	t	t	1320	Ö	ɔ́ɔ́	179
o	o	4156	Ê	ɛ́	1247	Õ	ɔɔ	167
m	m	4115	E	ɛ	1120	Ò	ɔ́ɔ	166
s	s	3679	ä	áá	909	æ	ee	157
j	j	3602	D	ʤ	894	Ë	ɛ́ɛ́	139
b	b	3550	Ô	ɔ́	760	ó	oó	137
k	k	3383	B	b^{v}	724	ù	úu	132
G	^{ŋ}g	3375	ð	^{n}ʤ	694	ö	óó	129
ô	ó	3227	T	ʦ	666	ü	úú	114
w	w	3111	á	aá	639	È	ɛ́ɛ	111
e	e	2713	W	^{m}w	621	u	uu	102
ê	é	2683	a	aa	585	Ó	ɔɔ́	95
n	n	2139	ñ	ɲ	534	í	ií	88
O	ɔ	1909	è	ée	528	g	ɣ	87
d	d	1853	ì	íi	397	Æ	ɛɛ	68
r	r	1846	ò	óo	357	É	ɛɛ́	55
Û	ú	1702	ï	íí	321	é	eé	49
à	áa	1622	õ	oo	303	ú	uú	38
ß	β	1597	ë	éé	299			

2.3. Dictionary

The transcriptions done by native Embosi speakers were phonemic in nature and were transcribed using IPA symbols. These symbols included strings of multiple characters representing both diacritics and phone symbols (e.g. 'mbv' represented a pre-nasalized bilabial trill). A list of all the phonemically transcribed words and their phonetic representations was made into a Latin-1 symbol to phone dictionary. As was already mentioned in section 2.2, the representation of Embosi consonants was fairly trivial, however, representing vowels in the Latin-1 phone dictionary was problematic.

Table 2: *Phonemic description of features for vowels*

Description and Feature	
$a1$	IPA symbol
$b1$	Latin-1 phone symbol
$[vowel/glide]$	vowel or semi-vowel
$[high/mid/low]$	vowel height
$[front/central/back]$	tongue position
$[short/long]$	vowel length
$[high/low]$	vowel tone
$[H/L/LH/HL]$	tone contour

We represented vowels with a number of features. Table 2 represents the mapping of IPA symbols ($a1$) to Latin-1 symbols ($b1$). Examples of actual mappings are given in Table 1. Vowels were additionally represented as the features shown in Table 2. Along with the vocalic features for tongue height and tongue position, suprasegmental features for vowel length, tone and tone contour were also used to create discrete vowel sets for each of the 7 Embosi vowels (i.e. /i, e, ɛ, a, ɔ, o, u/) which were represented in the ASR model used for the forced alignment segmentation of the speech corpus. These vowel features correspond to Clements's [11] method for a geometry of phonological features.

One drawback to the feature system used in Table 2 is that tone (and by extension tone contour) is treated as a phonemic feature of the vowel [12]. Tone is not represented as a distinct tonal tier separate from the vowel segment,[1] as it has been represented in many autosegmental model's of Bantu tone languages [14, 15, 16]. In consideration of vowel length, the phonological representation of tone is unclear for whether vowel length and tone act as independent phonemic features. This raises the following question: are two vowels represented as ÝV, such that the first vowel has a high tone and the second vowel has a low tone (represented with the absence of a tone accent), considered to be a sequence of a high vowel followed by a low vowel, or are the two vowels in the sequence considered to be a single long vowel having a contour falling tone (i.e. represented as a HL in Table 2)? Due to the architecture of the ASR model used here the exact nature of tone and vowel length is difficult to model. A 2-tier phonological representation of a speech segment is currently not supported by our ASR model. Thus, for our purpose of selecting Latin-1 phones for the ASR model, and for quantifying vowel elision, we considered that tone and contour (represented as HL and LH) are simply relevant distinctive features of the vowel in order to model the interaction of length and tone features in accordance to the phonological processes of vowel elision found in the Embosi speech corpus.

3. Results and Analysis

The investigation in this paper is focused on the interaction of morpheme deletion and vowel elision. The primary purpose of the investigation in this paper is to begin to quantify the vowel elision and morpheme deletion phenomena found in the Embosi speech corpus using ASR tools. Table 3 gives the frequency in the corpus of the following morphemes (*ya*, *mo*, *ba*, *ngá*, *nɔ*, *wa*) as they are deleted (n_{del}), as vowel elision occurs and they are deleted (n_{del+ve}), and as they occur with no deletion but possibly with vowel elision ($n_{\neg del}$). The total number of times

[1] See Goldsmiths's [13] distinction of tone being separate from a vowel, which is typically represented in autosegmental phonology as separate segmental and tonal tiers.

(N) that each morpheme was transcribed by the native speakers is also given. We considered native speaker's transcriptions to generally be phonemic, whereas the vowel elsions realized by our system are strictly phonetic. This is due to the nature of native Embosi transcribers being influenced by their knowledge of the language and its grammar. In contrast, our system does not make use of any high level grammatical information.

Table 3: *The total number of certain phonemically transcribed morphemes (N) in the Embosi speech corpus, and the frequencies of their deletion (n_{del}), their deletion with vowel elision (n_{del+ve}), and their occurrence (i.e. no deletion) but with possible vowel elision ($n_{\neg del}$) detected from the output of the ASR tools applied to the Embosi speech corpus*

Morpheme	n_{del}	n_{del+ve}	$n_{\neg del}$	Total N
ya	83 (35%)	125 (52%)	31 (13%)	239
mo	0 (0%)	0 (0%)	8 (100%)	8
ba	0 (0%)	0 (0%)	7 (100%)	7
ngá	12 (3%)	6 (1%)	439 (86%)	457
nɔ	13 (8%)	9 (6%)	133 (86%)	155
wa	17 (4%)	14 (3%)	431 (93%)	462

A simple observation from the data in Table 3 is that the frequency of morpheme deletion is not systematic. The morpheme *ya* has a considerably higher frequency of deletion than the other morphemes in the list. The grammatical function of this morpheme is the associative preposition. The morpheme *mo* is also a preposition, but it does not pattern at all as the morpheme *ya* patterns. Instead, its complete lack of deletion is similar to the class-noun nominal prefix morpheme *ba*. However, both of these morphemes have an extremely low total count overall in the corpus, such that it is difficult to make reliable judgments about them. The last three morphemes (*ngá*, *nɔ* and *wa*) are the singular subject pronouns. The frequency in which these morphemes delete is small. However, the results in table 3 do not discount the possibility that *vowel elision* may occur in further contexts than which are described in these results. The frequency in which they occur overall in the speech corpus is much greater than the other morphemes chosen in this investigation. It is possible that the low frequency of morpheme deletion is due to speech errors, quantifying errors or other reasons than a systematic reason in the phonology.

4. Discussion

As already mentioned, the central topic of this paper is on the quantification of vowel elision and morpheme deletion phenomena in Embosi, and how their quantification is represented in the Embosi speech corpus. In this section, we briefly describe the phonological processes that systematically apply across word boundaries in Embosi and the challenges in the development of a variant dictionary.

4.1. Vowel elision

To allow for dictionary variants that include the elision of vowels and deletion of morphemes as discussed in Rialland et al [17, 4, 18], variant rules in the dictionary were developed for vowel elision across word boundaries in cases of long/short vowel contact and in cases of short/short vowel contact. For the cases of long/short vowel contact (see (1)): vowel length, tone, and tone contour were necessary to accurately model the change in vowel length and change in vowel tone contour. For the case of short/short vowel contact (see (2)): vocalic features of tongue position and tongue height as well as tone were necessary to model the loss of a vowel, and the change in vowel tone.

(1) CV[$long, HL$]#V[$short$].CV $\rightarrow$ CV[$short, H$]#V.CV

(2) CV[H, low]#V[L, mid].CV $\rightarrow$ CV[H, mid]#CV

4.2. Morpheme deletion

Since morpheme deletion occurred, particularly in considering the results for the morpheme *ya* in Table 3, variants in the dictionary were created for the ASR model. In the phonemic transcriptions of sentences that contained the morpheme *ya*, an underscore character (i.e. '_') was added to the sentence to concatenate *ya* with the word immediately to its right in the sentence. This concatenated string was treated as a single lexical item, even though it was literally a string of 2 words, and this *new* lexical item was added to the dictionary. Several variant pronunciations were generated for each of these concatenated items in two contexts: (a) the concatenated left edge morpheme was deleted and only the right edge word was used to generate phonetic symbols to represent pronunciation variants; (b) the entire string of concatenated words/morphemes were used to generate phonetic symbols to represent pronunciation variants. Further considerations were also made to represent possible phonetic variations due to vowel elision internal to the concatenated string of morphemes, and to represent any possible phonetic variations due to vowel elision at the right and left edges of the concatenated strings.

The following possible conditions were represented as variants in the dictionary: (a) a condition in which no morpheme deletion occurred, but vowel elision did occur between the concatenated words and morphemes; (b) a condition in which morpheme deletion did occur, and vowel elision also occurred between the word that is immediately to the left of the concatenated string of morphemes and with the word that is at the right edge of the concatenated string. In this last condition, the deleted word is assumed to not be phonetically included for the purpose of vowel elision.

The conditions in which *ya* deletes are not due to phonetic or phonological principles, as is the case for vowel elision. Furthermore, the deletion of *ya* does not create any phonological boundary inhibiting vowel elision from occurring. This motivates our development of variant pronunciations in the dictionary where it is possible that either the morpheme *ya* is completely vacant from the utterance, or where it is also possible that the morpheme *ya* is present. The last consideration is for the fact that the deletion of *ya* does not occur in 100% of all instances that *ya* was phonemically transcribed in the speech corpus. In other words, *ya* does occur within certain contexts, particularly when at the beginning or at the end of an utterance.

4.3. Finding cases of vowel elision and morpheme deletion

The criterion for vowel elision used in judging the results from the ASR tool's forced alignment was when the system spent no more than 30 msecs on a vowel segment. This was the minimum time the system could spend on any given segment. These were the most likely cases where vowel elision occurred. The preceding word's vowel (from which the vowel elision occurred) was recorded. Each case was judged for whether vowel elision occurred according to the rule in (1) if the vowel contact was long/short, or according to the rule in (2) if the vowel contact was short/short. These judgments of whether vowel elision occurred or not were made automatically with analysis tools

written in Perl v5.10 that analyzed the output files of the ASR segmentation tools. Analysis tools were also developed to find cases where morphemes were deleted. This was a fairly trivial task since the ASR segmentation tool recorded the output of which variants were used from the pronunciation dictionary. These results were compared with the phonemic transcriptions to determine the frequency in which morphemes were deleted.

The ASR tool's segmentations were converted to textgrid formats that could be applied directly to corresponding spectrograms of audio files from the corpus [19, 20]. For example, in displaying the textgrid for sentence (3) shown in Figure 1, the Latin-1 phones were mapped to their corresponding IPA representations (cf. Table 1). The orthographic word level was also combined with the IPA mapping, and both are displayed under the utterance's spectrogram in Figure 1 to illustrate vowel elision and morpheme deletion detected by the system.

In this example, elision occurs between the vowel contact of the final vowel in the word *mwásí* and the initial vowel of the word *okondzi*. In order for the final vowel in *ya* to not have vowel contact with *okondzi*, *ya* must be deleted before vowel elision occurs. Possibly, this entails there is an ordering in the grammar of Embosi where deletion either occurs prior to vowel elision, or that *ya* deletes as a function of the morphology/syntax. The morpheme *ya* does not affect vowel elision at all in example (3).

(3) Wa ámitúbhá bɔ́si la mwásí ya okondzi
3S commit.PST adultery.14 with spouse.1 POSS Okondzi
'He committed adultery with Okondzi's wife.'[2]

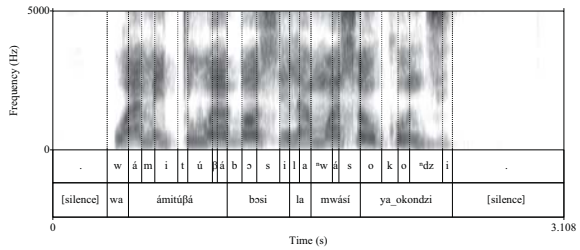

Figure 1: ya-*deletion and vowel elision*

It is a novel approach to use forced alignment and ASR tools in identifying occurrences of speech variants that directly relate to phonological and morphological processes. However, the reliability of the forced alignment output of the model needs to be considered. The current model used in this project has a high number of phones, but there is a low number of occurrences for many of the phones that are used in the training data (see Table 1). Several challenges to our approach should be mentioned. First, a large proportion of the phones used by the ASR tool are a result of our method for modeling tone, vowel length and the process of vowel elision in the language. However, vowel elision is more intricate than it is described here and as it has been implemented in the model. Vowel elision not only affects tone, but it also changes vowel quality across word boundaries. Second, the deletion of the morpheme *ya* has a distribution that is syntactically determined, but we only investigate this distribution in the context of omitting a small set of words. The syntactic context in which this distribution is determined is not directly considered in the model.

[2]The abbreviations used in the gloss are: 3S is third person singular; PST is the past tense; POSS is the possessive/associative preposition. Nouns are also specified by noun class numbers.

5. Conclusion

This paper has highlighted the initial findings of our study of the Embosi corpus, which is a part of the BULB project. The goal of the study so far has been to develop the ASR tools to perform forced alignment segmentation on the audio files in the speech corpus that have been prepared for study up to this point. Through these efforts we have identified phenomena noted in the literature regarding vowel elision. In the Embosi speech corpus vowel elision due to contact between long and short vowels has been observed, and vowel elision due to contact between short and short vowels has likewise been observed. Through the development of a variant pronunciation dictionary, we have been able to represent vowel elision using ASR tools.

We have also shown in this paper that there is also an interaction between vowel elision and morpheme deletion, specifically in cases where the morpheme *ya* is deleted. However, it is still unclear as exactly what mechanism causes *ya* to delete. We suggest the mechanism causing this is most likely not phonetic or phonological, but is morphological or syntactic instead.

6. Acknowledgements

This work was partly funded by the French ANR and the German DFG project BULB under grant ANR-14-CE35-0002, and the Labex EFL program under grant ANR-10-LABX-0083.

7. References

[1] G. Adda, S. Stukerb, M. Adda-Decker, O. Ambouroue, L. Besacier, D. Blachon, H. Bonneau-Maynard, P. Godard, F. Hamlaoui, D. Idiatov, G.-N. Kouarata, L. Lamel, E.-M. Makasso, A. Rialland, M. Van de Velde, F. Yvon, and S. Zerbian, "Breaking the unwritten language barrier: The BULB project," *Procedia Computer Science*, vol. 81, pp. 8–14, 2016.

[2] J.-L. Gauvain and L. Lamel, *Pattern Recognition in Speech and Language Processing*. CRC Press, 2003, ch. Large vocabulary speech recognition based on statistical methods, pp. 149–189.

[3] L. Lamel and J.-L. Gauvain, *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, 2005, ch. Speech recognition, pp. 305–322.

[4] A. Rialland, M. E. Amborobongui, M. Adda-Decker, and L. Lamel, "Phonologie et traitement automatique de la parole: le cas de l'embosi (bantu c25)," in *The American Conference of African Linguistics*, 2015.

[5] M. E. Amborobongui, "Processus segmentaux et tonals en mbondzi - (variété de la language ɛmbɨ́extopenosí c25)) -," Ph.D. dissertation, Universite Sorbonne Nouvelle - Paris 3, 2013.

[6] J.-M. Beltzung, A. Rialland, and M. E. Aborobongui, "Les relatives possessives en ɛmb´ɔsí," *ZAS Papers in Linguistics*, vol. 53, pp. 7–37, 2010.

[7] G. Adda and S. Stuker, "Breaking the unwritten language barrier (BULB)," ANR/DFG, Tech. Rep., 2014.

[8] E. Gauthier, D. Blachon, L. Besacier, G.-N. Kouarata, M. Adda-Decker, A. Rialland, G. Adda, and G. Bachman, "LIG-AIKUMA: a Mobile App to Collect Parallel Speech for Under-Resourced Language Studies," in *Interspeech 2016 (short demo paper)*, San-Francisco, France, Sep. 2016. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01350062

[9] L. Bouquiaux and J. Thomas, *Enquête et description des langues à tradition orale.Tome II: Approche linguistique (questionnaires grammaticaux et phrases)*., ser. SELAF. Peeters Publishers, 1976, vol. II, no. 230,1.

[10] R. P. Beapami, R. Chatfield, G. Kouarata, and A. Waldschmidt, *Dictionnaire Mbochi - Français*. Brazzaville: SIL-Congo, 2000.

[11] G. Clements, "The geometry of phonological features," *Phonology*, vol. 2, pp. 225–252, 1985.

[12] W. S.-Y. Wang, "The phonological features of tone," *International Journal of American Linguistics*, vol. 33, pp. 93–105, 1967.

[13] J. Goldsmith, "Autosegmental phonology," Doctoral dissertation, MIT, 1976.

[14] W. Leben, "Suprasegmental phonology," Doctoral dissertation, MIT, 1973.

[15] C. Paulian, *Le kukuya: Langue teke du Congo*. Paris: SELAF, 1975.

[16] L. M. Hyman, "Prosodic domains in Kukuya," *Natural Language and Linguistic Theory*, vol. 5, pp. 311–333, 1987.

[17] A. Rialland, M. E. Amborobongui, M. Adda-Decker, and L. Lamel, "Mbochi: corpus oral, traitement automatique et exploration phonologique," in *Traitement Automatique des Languages Africaines*, 2012, pp. 1–12.

[18] A. Rialland, M. Embanga Aborobongui, M. Adda-Decker, and L. Lamel, "Dropping of the class-prefix consonant, vowel elision and automatic phonological mining in embosi (Bantu C 25)," in *Selected Proceedings of the 44th Annual Conference on African Linguistics*, ser. Cascadilla Proceedings http://www.lingref.com/cpp/acal/44/index.html. Georgetown University: Ruth Kramer, Elizabeth C. Zsiga, and One Tlale Boyer, March 7-10, 2013 2015, pp. 221–230.

[19] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[20] P. Boersma and D. Weenink. (2013) Praat: doing phonetics by computer. [Online]. Available: Version 5.3.42: http://www.praat.org/