Social face to face communication – American English attitudinal prosody

Albert Rilliard [1], *Donna Erickson* [2,3], *Takaaki Shochi* [4], *João Antônio de Moraes* [5]

[1] LIMSI-CNRS, Orsay, France; [2] Showa University of Music, Kawasaki, Japan;
[3] Sofia University, Tokyo, Japan; [4] CLLE-ERRSàB, UMR 5263, Bordeaux, France;
[5] Laboratório de Fonética Acústica, FL/UFRJ/CNPq, Brazil

albert.rilliard@limsi.fr, ericksondonna2000@gmail.com,
takaaki.shochi@u-bordeaux3.fr, jamoraes3@gmail.com

10.21437/Interspeech.2013-427

Abstract

This paper presents the recording paradigm and the perceptual evaluation of a corpus of 16 prosodic social affects performed by a set of 8 native American English speakers (5f, 3m). The social affects are defined according to given communication goals in predefined social contexts, such as varying the relative hierarchical relation between the speaker and the interlocutor. The prosodic and facial strategies are evaluated by native listeners, rating their performance for achieving the targeted communication goal. Variations in the prosodic and facial strategies observed are then described and discussed in light of Ohala's frequency code. By selecting the best performances, 15 social affects were analyzed. Given the dimension of dominance as a main aspect related to the observed pitch level, the complexity of expressions is reflected in the multiparametric nature of the prosody. Voicing strength seems to be an important part of the acoustic information. In addition, the visual expressions allow an efficient interpretation of prosodic communication goals. Individual strategies to perform these social affects are observed in the prosodic variations, and may be related to factors such as the extraversion of speakers, their gender, and intrinsic pitch.

Index Terms: social affects, audio-visual prosody, face-to-face communication

1. Introduction

The prosodic expressions of social affects, or attitudes as defined by [1], are one of the means used by speakers to drive the illocutionary force of their intended speech acts [2] in face-to-face communication, together with e.g. lexical choice, word order or the verb's mood (cf. [3] for a discussion). Such strategic choices are partly linked with the speaker's own proficiency in the spoken language, her/his personality and the communication context (e.g. the hierarchical relation between interlocutors). These choices are also constrained at the linguistic level, individual languages having devoted specific formulae or conventionalized prosodic variations to specific contexts of interaction (cf. [4] for impoliteness strategies). Such kinds of expression are also supposed to rely on the proposed universal communication codes [5,6] – and chiefly the frequency code's relation with power during an interaction. From this code, one can hypothesize a lower fundamental frequency (F0) for the expressions of a dominant behavior (e.g. authority, arrogance), and a higher one for the more submissive behavior (e.g. politeness, interrogations).

In studying such kinds of prosodic variations that aim at targeted illocutionary acts during conversation, these different levels of variations are important. In the aim of foreign language teaching, the conventionalized ones were studied mostly [7,8,9]. In general, studies investigating such kinds of prosodic variations rely on stereotypic stimuli, and the corpus are based on the audio modality only [10,11,12,13]. Meanwhile, prosodic expressions rely on the audio-visual performance of speech [14], and it has been shown that visual cues are important for the decoding of such prosodic variations [15] – especially in the artificial situations encountered during most perceptual testing. Some works have shown that audio-visual presentation of such social affects clearly enhance understanding, especially in cross-cultural contexts [16, 17].

One common difficulty in the studies trying to compare the prosody of different kinds of social affects is linked to a difficult tradeoff between the high sound quality required for acoustic analysis, the need for a neutral lexical content of the studied sentences (ideally identical sentences for all the studied affects), the search for spontaneity of the expressions, and a clear labeling of the communicative goals of the speaker. Most of the cited studies use laboratory corpora. Typically, ad-hoc sentences are recorded by speakers trying to read a sentence and reproduce a given expressivity. To enhance the spontaneity of these expressions and to facilitate the speaker's task, [18] proposes to place target sentences in affectively-loaded texts; similarly, [13] recorded attitudinally-neutral sentences embedded into dialogues that prepare the speaker to perform an adequate expression for the target sentence.

The approach used during this research builds on these works. In order to study the expressive strategies used by speakers of varying linguistic backgrounds, communicative situations have been set-up so they can be plausibly used in different languages. They are described in the next section, while the corpus recording procedure is described in section 3. Section 4 presents the results of a performance evaluation test run on the obtained stimuli. These results are then discussed under the light of prosodic measurements of a sub selection of the best performances.

2. Social contexts for expression of attitudes

An important point of this research is to compare cross-cultural strategies for the expression of such social affects; enhancing foreign language teaching for such aspects of communication being one aim. Previous research on the cross-cultural variations in social affective expressions [8, 16] have been hampered by the use of labels to name these expressions, the translation of which raises problems well described by Wierzbicka [19, 20, 21]: there is no exact semantic correspondence between the terms used to define an attitude in two different languages, and that may lead to bias in the perceptual evaluations that are based on such translations.

To avoid this translation bias, communication contexts have been precisely described (the relative hierarchical levels of interlocutors, their social relation, the communication aim of the speakers), so two speakers may rely on this description to play a short dialogue that will lead to the production of target sentences. These contexts are described hereafter.

16 contexts have been selected, corresponding to a set of attitudes used in [8,16] for different languages. Some of these contexts don't have lexical equivalents in some languages, as the corresponding communication situations have not been conventionalized in a given culture. It is the case for example of the Japanese notion of *kyoshuku*, described by [22] as "corresponding to a mixture of suffering ashamedness and embarrassment, which comes from the speaker's consciousness of the fact his/her utterance of request imposes a burden to the hearer" (p. 34); *kyoshuku* has no lexical equivalent in e.g. English. Meanwhile, situations exist where an English speaker may feel something akin to *kyoshuku*.

The 16 social affects used in the present corpus are defined to speakers by the following prototypical situations (in each case, A is the recorded speaker, B the interlocutor):

- *Admiration (ADMI)*: A & B are almost the same age and know each other well. Both love French cuisine, and talk about the very delicious food they ate yesterday at a famous French restaurant. The scene is at a coffee shop.
- *Arrogance (ARRO)*: both A & B are from the same university, but A is older and A's father is head of the university and A is a bit of a snob. Both know each other, but are not friends. A organized a social party, and B was not invited to the party, but A is aware of his/her presence during the party. The scene is a party room, and A says to B that only his friends are invited.
- *Authority (AUTH)*: Speaker A is a custom agent; speaker B is a traveler. B is in front of A, requesting permission to enter the country; A needs to impose his authority; the scene is at a custom counter at the airport.
- *Contempt (CONT)*: both A & B are from the same university, but A is older; both know each other, but are not friends. In fact, A really hates B. A organized a social party, and speaker B was not invited, but A is aware of his/her presence. The scene is at a party room
- *Doubt (DOUB)*: A & B are colleagues, same age. A knows that his colleague B didn't go to the baseball game yesterday, but B pretends he went to the game, and A doesn't believe it. The scene is at a coffee shop.
- *Irony (IRON)*: A & B are friends, same age; A is going to Boston to see an important baseball game, and B, who is living in Boston calls A. Unfortunately, the weather in Boston is rainy and B says its wonderful; the scene is at an airport.
- *Irritation (IRRI)*: A & B are almost the same age and know each other. A is sitting next to B. Suddenly, B starts to smoke, and A is very angry; he wants him/her to stop, expressing his irritation toward speaker B. The scene is a public place.
- *Neutral declarative sentence (DECL)*: A & B are colleagues, same age; A gives information without any personal perspective; the scene is at a coffee shop.
- *Neutral question (QUES*: A & B are colleagues, same age. A asks for information, without any personal perspective, awaiting a simple answer. The scene is at a coffee shop.
- *Obviousness (OBVI)*: A & B are colleagues, same age; everyone knows B doesn't like French movies, but A asks B if he likes French movies or not; the scene is at a coffee shop.
- *Politeness (POLI)*: A & B are almost the same age and don't know each other well, but work together professionally. A is sitting next to B; both start social talk. The scene is at a formal party.
- *Seduction (SEDU)*: A loves B and they have an intimate relationship. A gives a compliment to B in a sexually provocative way. The scene is at a club house.
- *Sincerity (SINC)*: B is chief of the section which A belongs to; B is older than A. The chief (B) wants A to take on a big project; A is pleased to be asked to do this, and expresses his enthusiasm, honesty and sincerity for this task. The scene is at B's office.
- *Surprise (SURP)*: A & B are friends, same age. A didn't know that B can sing well. One day, B makes A listen to his beautiful voice. The scene is at friend's home.
- *Uncertainty (UNCE)*: A & B are colleagues, same age. A saw B at the baseball game yesterday, but is not 100% sure if it was really B; the scene is at a coffee shop.
- "*Walking on eggs*" (WOEG): B is chief of the section which A belongs to; B is older than A. The chief (B) wants A to do a task which is a lot of work, and it seems to A it is impossible to do this, so A tries to reject this request by trying to make sure her/his boss (B) doesn't get angry for refusing. The scene is at B's office.

The last social affect, "walking on eggs", corresponds to a certain extent to the Japanese *kyoshuku* situation, adapted to an American English situation.

For all situations, two short neutral target sentences have been used to record the respective prosodic expressions: "*Mary was dancing*." and "*A banana*." In order to induce these target sentences in each context, small dialogues were written (cf. [13]), that take place in the prototypical context described above, and that end with one of these sentences. Dialogues for the "*banana*" sentence are mostly based on pictures such as the one found in figure 1, for doubt, while written dialogues are used for the other sentence. Currently, these situations have been adapted to three languages: American English, Japanese and French. The present paper focuses on the American English results.

Figure 1: *picture used during the recording to induce doubt with the "*banana*" sentence.*

3. Corpus recording procedure

Most speakers were recruited amongst university students, and were paid for their performance. As one aim of the corpus is to address the cross-cultural social affect specificities for foreign language teaching, all speakers are first recorded in their second language, and then in their native language. The native speakers of American English (5 females, 3 males), whose results are reported here, were recruited in Japan, and first recorded in Japanese, then in American English. During the recordings, each speaker (A in the above situations) has an active interlocutor (B in the above situations) who interacts with her/him – in order to enhance the natural of the communication situation, and to ease the production of realistic expressions. Speaker B is native speaker, and generally the same speaker for all the recordings.

The recordings took place in a sound-treated room. The speaker A (head and chest visible) was recorded on a Panasonic AG-AC160 video camera recording video in AVCHD format, with the sound of both speakers captured by a Earthworks QTC1 omnidirectional microphone, placed at one meter from speaker's A mouth (this distance was chosen to limit the influence of the speaker's movements on the sound level). The microphone level was calibrated before each recording session using a Brüel & Kjær acoustical calibrator, thus the sound pressure level can be corrected after recording to a level comparable across all speakers.

The target sentences were then manually searched for across the corpus, isolated and extracted into individual video files. Any speech utterances from speaker B occurring during the expressive gesture of speaker A performing one target sentence were removed from the sound track (none were overlapping with their speech). Due to the interactive nature of the recording, some spontaneous changes were observed on the target sentences: typically, the "banana" sentence may be performed as "a banana", "banana", "it's a banana", while "Mary was dancing" was sometimes performed as "She was dancing". Speakers also use sometimes interjections, such as "hmm", "er", "oh", etc., together with the target sentences. The wave of each stimulus was hand-labeled at a phonetic level using the PRAAT software [23].

4. Performance evaluation

4.1. Subjects & Procedure

17 subjects (7 females, mean age 25), all native American English speakers, listened to 256 stimuli (8 speakers performing 16 attitudes with two sentences) and rated the performance of the speaker in expressing the targeted attitude, on a 1 to 9 scale. Stimuli presentation is grouped by speaker, so listeners may concentrate on the specific strategies used by each speaker, and not on a comparison of the different speakers. Stimuli are presented in their audio-visual performance. Due to the number of stimuli, the experimental procedure is automatized to a large extent. Prior to the experiment, subjects are presented with the 16 social affective labels and their contextual descriptions. Then, before each stimuli, subjects are presented with the targeted social affect during 2.5 seconds; the stimuli is then played; the subject has then 10 seconds to rate the performance for the given social affect using the keyboard from 1 to 9. A slider indicates him how much time remains. After the answer (or the 10 seconds), the next stimulus is automatically launched. Only 7 answers were not given in the 10 second allotted response time. Subjects can pause between each speaker.

4.2. Results analysis

Results for the 17 subjects were pooled and an ANOVA (completely randomized three-factorial design) was run using R software [24]. The answers given by each individual listener were standardized calculating their z-score value, in order to remove any individual tendency to use part of the proposed evaluation scale. The observed variable is thus a standardized z-score. The three factors were the Speaker (Sp, 8 levels) that produces the stimulus; the Attitude (At, 16 levels) which was intended; the target Sentence (Se, 2 levels) used to produce the attitude. The significance level was set at 0.01. Results of the ANOVA are presented in table 1. A significant effect of the speaker and attitude was measured on the performance, while the sentence's effect was not significant at the 1% level. The interaction between speaker and attitude is also significant.

Table 1. *ANOVA table for the performance evaluation.*

	df	df error	F	p	Partial η^2
Speaker	7	4096	25.54	0.000	0.042
Attitude	15	4096	23.54	0.000	0.079
Sentence	1	4096	4.74	0.029	0.001
Sp*At	105	4096	2.94	0.000	0.070
Sp*Se	7	4096	2.83	0.006	0.004
At*Se	15	4096	1.46	0.111	0.005
Sp * At * Se	105	4096	1.86	0.000	0.045

Figure 2 show the relative performances of the 8 speakers, in decreasing order (note that results are not reported in z-score, as such data is less straightforward to interpret). A post-hoc Tukey test shows that the one female speaker globally performs significantly better than the other four female speakers. However, the decrease in performance from level 7 to 6 is fairly linear and there is no significant difference in one speaker to the next.

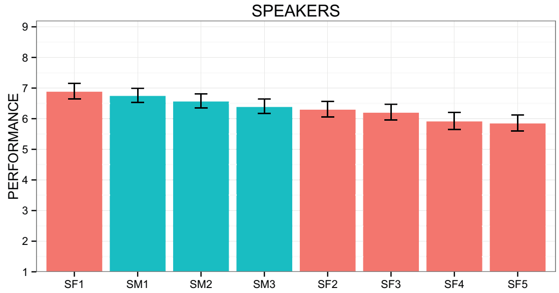

Figure 2: *mean performance of each speaker, F is for female and M for male speakers.*

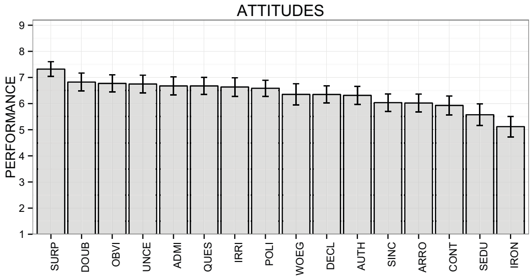

Figure 3: *mean performance of each attitude (see text for labels).*

Figure 3 presents the relative performance level obtained by each attitude. Surprise is the attitude the most consistently highly rated across speakers, although not rated significantly better than the next 6 best performances. Irony, rated at the middle of the scale, received significantly lower scores than all others attitudes, except seduction (post-hoc Tukey test). Both irony and seduction are two expressions that are highly contextually dependent, and for which individual variants may have great importance. For the remaining analysis, (prior to a more detailed analysis of irony's performances), this expression will not be further analyzed. For the 15 remaining attitudes, their prosodic parameters have been estimated and hand corrected: F0 in semitones, intensity in dB (these two parameters are calculated on the vowels only and normalized for each speaker's mean), z-score of the phonemic duration using the method in [25].

The significant interaction between speakers and attitudes highlight the differences that individual performances may have on the perception of such social affects. In order to pay attention to detailed prosodic variations, a sub selection of the best performances, for each attitude and target sentence, has been done, and these sentences' audio realizations have been carefully observed. The next section will discuss the most prominent findings of these acoustic analyses; due to space restriction, the "banana" sentence only is analyzed here.

5. Discussion & Conclusion

Table 2 presents the mean and standard deviation for the prosodic parameters observed on the "banana" sentence, for each attitude, all speakers averaged.

Table 2. *mean (m) and standard deviation (sd) for the F0, intensity, and z-scores of duration.*

ATTIT	F0 (ST)		Intensity		z-duration	
	m	sd	m	sd	m	sd
CONT	-2,68	3,44	1,18	5,59	-0.324	0.720
ARRO	-2,21	3,62	2,52	5,48	-0.228	0.694
AUTH	-2,23	4,27	2,90	4,84	-0.453	0.565
SEDU	-2,80	3,39	-2,74	4,30	0.260	1.003
DECL	-2,02	2,82	1,91	5,33	-0.478	0.518
OBVI	-1,68	4,40	1,65	5,54	0.042	0.957
IRRI	1,27	3,07	7,76	4,69	0.318	1.073
ADMI	6,69	4,76	9,92	6,38	0.669	1.361
POLI	0,78	3,51	0,26	4,60	-0.312	0.670
SINC	-0,88	3,78	-0,11	5,02	-0.110	0.833
WOEG	-1,46	2,91	-1,01	4,63	0.419	1.144
UNCE	0,11	4,02	-0,53	3,88	0.377	1.190
QUES	1,01	5,01	1,17	3,22	-0.384	0.813
DOUB	0,60	7,12	0,84	5,17	0.211	1.082
SURP	3,10	6,04	5,69	4,48	-0.052	0.961

As stated in the introduction, the frequency code suggests that dominant behaviors may be accompanied by strategies for the speaker to appear larger and thus more powerful – hence a lower F0; and conversely, more submissive or interrogative behavior may be performed with a higher voice. This tendency is strongly supported by our results, where contempt, arrogance, authority (all expressions where the speaker looks down on the interlocutors or aims at asserting power) are performed with a clearly lower F0 than politeness, sincerity, uncertainty, question, doubt, surprise (all expressions in which the speaker is not assuming a position of dominance over the interlocutor). Other expressions where the speaker imposes his will or looks down on the interlocutor, however, don't strictly follow this scheme. For instance, the expression of irritation, presumably a situation in which the speaker asserts power, has a relatively high F0 but note that it is has very high intensity, compared to other dominant expressions; it may be the increased lung pressure causing the increased F0. Irritation also departs from the other dominant expressions due to clear lengthening observed on some phonemes (cf. standard deviation of duration).

The extremely high F0 and intensity used for admiration link it with an expression of activated joy, in this case, the idea of receiving something very pleasant (in the corpus framework, a splendid banana). It is accompanied by lengthening on the stressed and final syllables. ”Walking on eggs", which corresponds to a complex situation, and is not conventionalized in the American English language, expresses a mixture of suffering and shame, but the speaker also imposes something (i.e., the truth) on his interlocutor. The speaker's position is clearly inferior to the interlocutor's. It seems that the low F0 is linked to low intensity, perhaps indicative of the speaker's feeling of lack of power. Measures of voice source quality may improve the description of this affect.

Interestingly, the expressions of seduction and declaration are also performed with a low F0, although they are not typically expressions of dominance. The expression of obviousness shows values in a middle F0 range that may correspond to its rather subtle meaning.

In order to better understand what might be some of the acoustic cues for contempt, arrogance, and authority, we examined those "banana" sentences which were produced by the 5 speakers who had the highest performance scores. All speakers showed almost the same linearly decreasing F0 contour, with a low F0 range (which is coherent with the mean measures shown in Table 2). However, careful examination of individual speaker strategies suggested some interesting phonetic differences among these three affective expressions. For *Contempt*, three speakers had a tongue click or short expiration before "banana" (as if to indicate "distaste" directed at the interlocutor); two speakers showed a slight F0 continuation rise on the final vowel; three speakers produced the stressed (middle) vowel more tense, by either fronting the vowel (higher second formant) or tensing the vocal folds (higher spectral tilt); one speaker produced contempt with the fastest speed, the shortest final vowel and the least force of articulation.

For *Authority*, two speakers showed reduced fourth formants (which may be related to a lowered larynx, hence a lengthened vocal tract, presumably to represent a larger, more dominant person, as predicted by the frequency code); two speakers had increased first formant (thus opened mouths more); and one speaker had stronger articulation, especially the /b/ of "banana".

For *Arrogance*, one speaker changed the stressed /ae/ vowel to a more "snob-like" /a/ vowel; one speaker spoke more slowly compared with authority and contempt.

These results generally support the frequency code, but also show the importance of fine phonetic details for understanding the distinctions among social affective expressions. More detailed acoustic analysis will be performed in order to have a better understanding of the rich prosodic variability at hand.

6. Acknowledgements

This work was supported by an ANR grant PADE. The authors warmly thank the speakers as well as Mariko Kondo and Sylvain Detey from Waseda University for their help.

7. References

[1] Wichmann, A., “The attitudinal effects of prosody, and how they relate to emotion”, in Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, 143–148, 2000.

[2] Fónagy, I., Bérard, E. and Fónagy, J., “Clichés mélodiques”, Folia Linguistica 17:153–185, 1984.

[3] de Moraes, J.A. & Rilliard, A., “Illocution, Attitudes and Prosody”, in T. Raso et al., Spoken Corpora and Linguistic Studies, John Benjamins, to appear.

[4] Culpeper, J. Bousfield, D. and Wichmann, A., “Impoliteness revisited: with special reference to dynamic and prosodic aspects”. Journal of Pragmatics, 35:1545-1579, 2003.

[5] Ohala, J.J., “An ethological perspective on common cross-language utilization of F0 of voice”, Phonetica, 41:1-16, 1984.

[6] Gussenhoven, C., “Intonation and interpretation: phonetics and phonology”, in Proceedings of Speech Prosody 2002, Aix-en-Provence, 2002.

[7] Martins-Baltar M., “De l’énoncé à l’énonciation: une approche des fonctions intonatives”, Paris: Didier, 1977.

[8] Shochi, T., Rilliard, A., Aubergé, V. & Erickson, D., “Intercultural perception of English, French and Japanese social affective prosody”, in S. Hancil (ed.), The role of prosody in affective speech, Linguistic Insights 97, Bern: Peter Lang, AG, Bern, 31-59, 2009.

[9] Lu, Y., Aubergé, A. & Rilliard, A., “Do you hear my attitude? Prosodic perception of social affects in Mandarin”, In Proceedings of the 6th International Conference on Speech Prosody (SP 2012), Shanghai, China, 685-688, 2012.

[10] Fujisaki, H. & Hirose, K., “Analysis and perception of intonation expressing paralinguistic information in spoken Japanese”, Proceedings of the ESCA Workshop on Prosody, 254-257, Lund, Sweden, 1993.

[11] Morlec, Y., Bailly, G. & Aubergé, V., “Generating prosodic attitudes in French: Data, model and evaluation”, Speech Communication, 33(4):357–371, 2001.

[12] de Moraes, J. A., “The pitch accents in Brazilian Portuguese: Analysis by synthesis”, in Proceedings of Speech Prosody 2008, Campinas, 389–397, 2008.

[13] Gu, W., Zhang, T. & Fujisaki, H., “Prosodic Analysis and Perception of Mandarin Utterances Conveying Attitudes”, Proceedings of Interspeech, Firenze, Italy, 1069-1072, 2011.

[14] Swerts, M., & Krahmer, E. 2005, “Audiovisual prosody and feeling of knowing”, Journal of Memory and Language 53(1):81-94, 2005.

[15] Nadeu, M. & Prieto, P., “Pitch range, gestural information, and perceived politeness in Catalan”, Journal of Pragmatics, 43(3): 841-854, 2011.

[16] Rilliard, A., Shochi, T., Martin J.C., Erickson D. & Aubergé, V., “Multimodal Indices To Japanese And French Prosodically Expressed Social Affects”, Language & Speech, 52(2/3):223-243, 2009.

[17] de Moraes, J. A., Rilliard A., Mota B. & Shochi T., “Multimodal perception and production of attitudinal meaning in Brazilian Portuguese”, in Proceedings of Speech Prosody 2010, Chicago, paper 340, 2010.

[18] Grichkovtsova, I., Morel, M., & Lacheret, A., “The role of voice quality and prosodic contour in affective speech perception”, Speech Communication, 54(3):414-429, 2012.

[19] Wierzbicka, A., “A semantic metalanguage for a cross-cultural comparison of speech acts and speech genres”, Language in Society 14(4): 491-513, 1985.

[20] Wierzbicka, A., “Defining Emotion Concepts”, Cognitive Science 16:539-581,1992.

[21] Wierzbicka, A., “Empirical Universals of Language as a Basis for the Study of Other Human Universals and as a Tool for Exploring Cross-Cultural Differences”, Ethos 33(2):256–291, 2005.

[22] Sadanobu, T., “A natural history of Japanese pressed voice”, Journal of the Phonetic Society of Japan 8(1): 29-44, 2004.

[23] Boersma, P. & Weenink, D., “Praat: doing phonetics by computer [Computer program]. Version 5.3.32 retrieved 17 October 2012 from http://www.praat.org/

[24] R Core Team, “R: A language and environment for statistical computing”. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/, 2012.

[25] Campbell, N., “Automatic detection of prosodic boundaries in speech”, Speech Communication, 13:343-354, 1993.