Creation of a Dialogue Corpus for Automatic Analysis of Phonetic Convergence

Jolanta Bachan, Mariusz Owsianny, Grażyna Demenko

Institute of Linguistics,
Adam Mickiewicz University in Poznań, Poland
jolabachan@gmail.com, mowsianny@man.poznan.pl, lin@amu.edu.pl

Abstract

The current work presents the creation of a dialogue corpus for analysis and formal evaluation of phonetic convergence in spoken dialogues in human-human and human-machine communication, with the goal of comparing dialogue features at all levels of language use. The corpus was created within an ongoing project which aims at (1) extracting phonetic features which can be mapped on a synthetic signal, (2) creating dialogue models applicable in a human-machine interaction and (3) practical evaluation of the convergence. For the corpus, 16 pairs of Polish speakers were recorded in a professional studio. The speakers could hear each other over the headphones, but could not see each other. For the recordings, 2 overhead microphones and 2 stationary microphones were used, providing 4 mono channels of recordings. The recording scenarios consist of controlled, neutral and expressive tasks. This scenario combination was novel and promises to provide an empirical foundation for both linguistic and computational dialogue modelling of both face-to-face and man-machine dialogue.

Keywords: dialogue corpus, phonetic convergence, recording scenarios, human-computer interaction

1. Introduction

Phonetic convergence in a dialogue is a natural phenomenon. Phonetic convergence involves shifts of segmental as well as suprasegmental features in pronunciation towards those of a communicative partner (Pardo 2006). The research on this phenomenon has its origin in the Communication Accommodation Theory (CAT) that has been established in the 1970s (Giles 1973, Giles et al. 1991). The main assumption of this theory is that interpersonal conversation is a dynamic adaptive exchange involving both linguistic and nonverbal behaviour between two human interlocutors. This theory started as a model of interpersonal communication and has since been developed to encompass insights from a number of disciplines, including linguistics, sociology and psychology. One central ingredient of CAT is the attention that speakers and listeners direct at the speech of their interlocutors. Individual adjustments to speech are assumed to subserve the function of controlling (maintaining, reducing or increasing) social distance. The speaking style of conversational partners thus converges, diverges or remains unchanged, depending on the strategies applied by the interlocutors. Most studies in the CAT framework aim at finding social motivations for accommodation behaviour and share the assumption that the processes underlying the manipulation of speech behaviour are – at least partially – under the speaker's conscious control.

Speakers accommodate their behaviour on semantic, lexical, syntactic, prosodic, gestural, postural and turn-taking levels (Pickering & Garrod 2004). The function of inter-speaker accommodation is to support predictability, intelligibility and efficiency of communication, to achieve solidarity with, or dissociation from, a partner and to control social impressions. The significant role of such adaptive behaviour in spoken dialogues in human-to-human communication has important implications for human-computer interaction. In the context of speech technology applications, communication accommodation is important for a variety of reasons: models of convergence can be used to improve the naturalness of synthesised speech (e.g. in the context of spoken dialogue systems, SDS), accounting for accommodation can improve the prediction of user expectations and user satisfaction/frustration in real time (in on-line monitoring) and is essential in establishing a more sophisticated interaction management strategy in SDS applications to improve the efficiency of human-machine interaction.

Studies on phonetic convergence rest on the assumption that the incoming speech signal undergoes an early, front-end analysis, which decomposes the speech signal into a set of features. In principle, each feature can be the target of convergence processes in production. Acoustic features investigated include (e.g. Ward & Litman 2007, Baumann & Grice 2006, Gorisch et al. 2012): voice-onset time (VOT), formants, voicing, F0 range and register, pitch accents, intensity, duration, pausing, and speaking rate, as well as the long-term average spectrum (LTAS). Such acoustic measures can be complemented by perceptual judgements of the presence or degree of convergence.

Communicative adaptation has been viewed as a potential functionality in human-machine interaction to improve system performance (Edlund et al. 2008, Edlund et al. 2006, Carlson et al. 2006, Porzel & Baudis 2004, Porzel et al. 2006, Bachan 2011a). It can be assumed that a responsive human-computer interface that accommodates some features of the human interlocutor may be perceived as more user-friendly and may even lead to enhanced learning. The phenomenon of phonetic convergence that occurs naturally and partly automatically in human-human communication has not yet been exploited sufficiently in human-machine communication systems and the manipulation of the phonetic structure of speech generated in SDS environment with the aim of converging to the human's speech pattern has been hardly investigated so far (cf. Bachan 2011b, Lelong & Bailly 2011, Savino et al. 2016, Oertel et al. 2016).

Apart from being an information exchange, it is widely recognised that human conversation also is a social activity that is inherently reinforcing. As such, new conversational interfaces are considered social interfaces, and when we participate in them we respond to the computer linguistically and behaviourally as a social partner. Human-computer interfaces that mimic human communication (and thus account for accommodation/convergence phenomena) will constitute next-generation conversational interfaces for speech technology applications. The benefits of using speech as an interface are multiple: simplicity (speech is the basic means of communication), quickness, robustness, pleasantness (related to social aspects of spoken communication, building relations), convenience (it can be used in hands-free and eyes-free situations or when other interfaces are inconvenient), it can be used as an alternative interface for the disabled and has some technical benefits (readily available hardware such as a telephone is sufficient).

Although the literature on communication accommodation in spoken dialogues in human interaction is fairly extensive, research on human-computer interaction has yet to face the challenge of investigating whether users of a conversational interface likewise adapt their speech systematically to converge with a computer software interlocutor. At this moment, the application of phonetic convergence in speech technology applications is not feasible for two reasons. The first one is related to the lack of an efficient quantitative description of this complex behavioural phenomenon as it occurs in spoken language. Past research on interpersonal accommodation has focused on qualitative descriptions of the social dynamics and context involved in linguistic accommodation. It also has relied on global correlation measures to demonstrate linguistic accommodation between two interlocutors. Only quantitative predictive models that account for the magnitude and rate of adaptation of different features, the factors that drive dynamic adaptation and re-adaptation, and other key issues will be valuable in guiding the design of future conversational interfaces and their adaptive processing capabilities. The second reason is that current SDS architectures are not designed to accommodate natural dialogue with human users, therefore a platform for testing quantitative models of inter-speaker accommodation does not yet exist.

The present paper describes spoken dialogue corpus creation for analysis and objective evaluation of phonetic convergence in human-human communication. The analysis of the corpus will serve for creation of convergence models which could be implemented in spoken dialogue systems based on spontaneous, expressive speech.

2. Corpus design

The corpus is being created within an ongoing project which aims at (1) extracting phonetic features which can be mapped on a synthetic signal, (2) creating dialogue models applicable in human-machine interaction and (3) practical evaluation of the types and degree of phonetic convergence. It is planned to record dialogues with different configuration of speakers' L1 / L2:

- Polish L1 speaker with Polish L1 speaker
- Polish L1 speaker with German L1 / Polish L2 speaker
- German L1 speaker with German L1 speaker
- German L1 speaker with Polish L1 / German L2 speaker

The present paper describes only the creation of the dialogue corpus between the Polish L1 speakers. The recordings of the dialogues between the other groups is planned as the next step of phonetic convergence analysis, also in different language pairs.

2.1. Subjects

For the corpus, 16 pairs of Polish speakers were recorded: 8 male-male pairs and 8 female-female pairs who knew each other and/or were close friends. From all the subjects such metadata was collected as: name, sex, age, height, weight, education, profession, information on languages spoken and proficiency levels.

The youngest subject was 19 years old and the oldest was 58 years old (recorded in pair with a 50-year-old), the biggest age difference was 12 years and the average age difference was 3 years. Only 3 pairs of female speakers were above 30 years old, all the other subjects were younger than 29 years. The average age of the subjects was 27 years. Additionally, in each session a 33-year-old female teacher/phonetician carried out 3 dialogues with each of the subjects.

2.2. Scenarios

The recording session was composed of a few short tasks. The first tasks were controlled reading and repetition. These tasks were introduced to assess the speakers's talent to adapt their speech to the model voice and their expressiveness while reading an enthusiastic interview with a music star. The next set of dialogues were task-oriented (neutral): either the dialogues were cooperative with no leader or in the dialogue the leader was specified and it was expected that the interlocutor would adapt to the leader's voice. Additionally, a set to expressive scenarios was recorded. These dialogues were also cooperative with no leader in the dialogue when recorded in pairs of common speaker, but when each of the speakers was to talk with the teacher/phonetician, it was expected that the speaker would adapt to the teacher in their expressiveness, liveliness and language.

Such a choice of scenarios was made to apply the developed convergence models to speech technology scenarios at different kinds of call centers, automatic information services or computer games.

2.2.1. Controlled scenarios

There were 3 tasks in the controlled scenarios. In the first task, the subject heard a recording of a short sentence over the headphones by a male or a female speaker and the subject's task was to repeat the sentence in a way to best imitate the melody of the original. The sentence "Jola lubi lody" (Eng. "Jola likes ice-creams") was played 6 times with a stress on different syllables: "**Jo**la lubi lody" or "Jola **lu**bi lody" or "Jola lubi **lo**dy". Figure 1 shows the short sentence uttered by the male or female speaker, with the stress marked by "+" on different syllables.

The second task was to read a dialogue. It was an interview by a reporter and a singer. The dialogue was constructed in such a way to contain neutral and expressive phrases with exclamations.

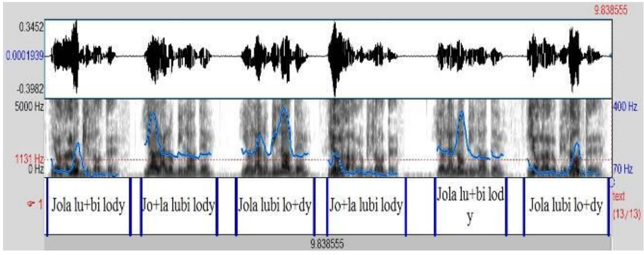

Figure 1: The sentence "Jola lubi lody" with stress marked by "+" on different syllables.

In the third task the subject was to read/repeat the phrases of the same dialogue, but imitating the melody of phrases of the pre-recorded speech (a similar task as in the first one, but this time the sentences were longer and their expressiveness differed).

These controlled recordings were carried out to evaluate general speakers possibilities to produce segmental and suprasegmental structures (accent type and placement, consonant cluster production) and to assess whether the speakers had talent to imitate other's speech and whether they could be expected to phonetically converge with the other speaker to a great extent. While recording the corpus, two phoneticians carrying out the recordings assessed perceptually that one speaker had little tendency to adjust his speech to the speech recordings.

2.2.2. Task-oriented scenarios

The task-oriented (neutral) scenarios consisted of 4 dialogues. The first was a decision-making dialogue in which the interlocutors were to decide together what to take to a desert island to survive. They could choose 5 items from the following list: TV set, binoculars, matches, nails, soap, favourite teddy bear, mattress, knife, petrol, tent, pen, bowl, book, hammer, kite. This was a cooperative dialogue, there was to be no role asymmetry and the maximum convergence was expected.

The second dialogue was based on a diapix task (van Engen et al. 2010) where in a cooperative dialogue the subjects were to find 3 differences between two pictures. There was no role asymmetry and the subjects had to describe their pictures in order to find the differences. The diapixes are presented in Figure 2. There are 10 differences between the pictures, but preliminary recordings revealed that finding all differences was taking too long and the task was simplified to finding only 3 differences.

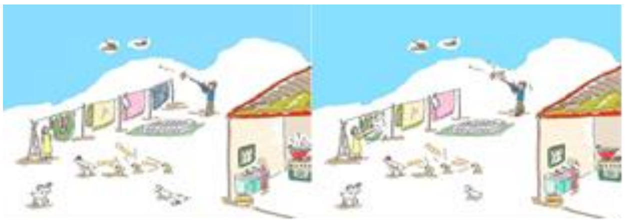

Figure 2: Diapixes for neutral scenario: describe and find 3 differences (van Engen et al. 2010).

The last two dialogues from the task-oriented scenarios were map-tasks. One of the speakers was asked to play a tourist in a foreign city who just arrived at the main station and the other was to pretend to be a receptionist in a hotel. The tourist was calling the hotel at which he booked a room to ask how to get there from the main station. The subjects had the map of the city to be used in the dialogue (Figure 3). There was asymetry in the dialogue and it was expected that the tourist would converge to the receptionist, i.e. the leader of the dialogue. The map-task was recorded twice with the speakers exchanging their roles and with different maps.

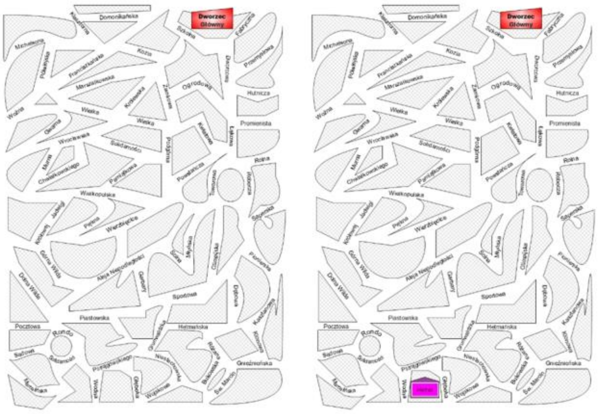

Figure 3: Maps for the map-task: tourist's map on the left, receptionist'**s** map on the right; "Dworzec Główny" means "Main Station."

2.2.3. Expressive scenarios

The set of expressive dialogues was divided into 4 groups: a) asymmetry: power – dominant vs. submissive (entertainment scenario), b) asymmetry: emotionally coloured speech – valence: positive vs. negative (fun vs. sadness/fear, terrorist attack scenario), c) no role asymmetry: both speakers in agreement vs. both speakers in disagreement (provocation in art) and d) dialogues with the teacher (also agreement and disagreement).

In the first scenario one of the speakers played the role of a tourist information centre assistant of a big city and his task was to provide information about events and interesting places in the city and to convince the caller to choose at least one of his offers. If he had convinced the caller, the assistant would have received an award from his boss. The other person was a party-goer who wanted to find out what attractions the city offered at night. The dialogue was asymmetric, designed to boost a strong convergence to the tourist information assistant, the leader of the dialogue, who showed great enthusiasm. The same scenario was used again, but with the exchanged speakers roles.

In the second scenario, the tourist information assistant was informed about terrorist attacks in the city and was unwilling to provide any information about the entertaining events in the city. Despite the threat of another attack, the assistant has to inform the caller about the interesting places in the city, but the best procedure was to suggest only the safest options or to convince the caller to stay at home. The other speaker was again the party-goer who despite the threat of terrorist attacks wanted to go out to have some fun. The dialogue was to show a strong asymmetry and convergence to the assistant, the leader, who showed no enthusiasm to provide any information and even scared the caller that going out might put his life in danger. After the dialogue was finished, the subjects changed their roles and carried out a similar dialogue again.

Dialogues on provocation in art were designed to elicit mutual convergence as there was to be no role asymmetry. The subjects saw pictures of a very provocative content and their tasks were to discuss them and approve this form of art in the first scenario, and later they both were asked oppose and condemn such art. The same set of approve/oppose dialogues was also carried out between each subject and the teacher. Although the dialogues in this part of the corpus are assessed as the most interesting by the creators, for the future less historically sensitive topic is to be chosen.

Finally, the last dialogue between the teacher and the subject was about Madonna's provocative performance. Both parties strongly supported their own views. The teacher – the opponent – was very conservative and thought Madonna was evil and condemned Madonna for crucifying herself during her concert. Contrary, the subject – the supporter – was a fan of modern art, liked provocations and loved Madonna. Their task was to exchange their opinions of the presented photo from Madonna's concert (Figure 4).

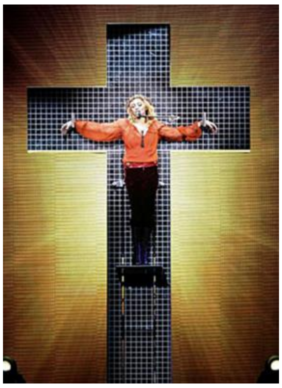

Figure 4: Picture for the expressive scenario: Madonna on the cross.

The dialogues with the teacher allowed to control the dialogues, boost more expressiveness if needed, more fun or extreme indignation. The teacher could also control the length of the dialogues and make it longer if she thought the given subject did not speak long enough.

2.3. Recording session

The recordings were carried out in a professional recording studio in the Institute of Linguistics at Adam Mickiewicz University in Poznań in 2016. The recording session started by signing an agreement by the subjects for recording their voices in audial carriers, for using the recordings for analyses and for the possibility of multiplying the recorded voices by using any technique in any amount in every known carriers. Speakers answered also the questions concerning basic personal information described in section 2.1 Subjects.

For the dialogue recording, the studio has been specially prepared according to the highest standards (Gibbon et al. 1997) – participants of the experiment felt free and could hear each other over the headphones but could not see one another. One of the recorded persons was closed in the insulated reverberation cabin while the second speaker sat in the corner of the studio which was separated by sound absorbing panels. The speech prompts were on a piece of paper, but during the recording the speakers were asked to put the paper on a small table nearby. Holding a paper is a classic source of noise and for the future recordings a music stand will be used during the recording sessions.

Four professional microphones were used for recordings: 2 overhead microphones (DPA 4066 omnidirectional headset microphone) and 2 stationary microphones (condenser, large diaphragm studio microphone with cardioid characteristic – Neumann TLM 103). Microphones were plugged into the high performance audio interface Roland Studio Capture USB 2.0 equipped with 12 microphone preamps. The recordings were carried out using Cakewalk Sonar X1 LE software. This setup provided 4 mono channels of recordings, 2 for each speaker, at 44.1 kHz sampling frequency and 16 bit depth. Exemplary screenshot showing the process of recording a dialogue is presented in Figure 5. First speaker's voice was recorded in sound insulation cabin (anechoic chamber): first sound track is recorded using studio stationary microphone and the third sound track was recorded with the headset microphone. Second and fourth sound tracks concern respectively studio and headset microphones used by second speaker in the acoustically separated by sound absorbing panels corner of the studio. One recording session lasted approximately 2 hours and provided about 1 hour of speech. During the recordings, the speakers were asked to drink mineral water to refresh their throats. Short breaks were also taken if needed. Altogether over 13 hours of speech was collected.

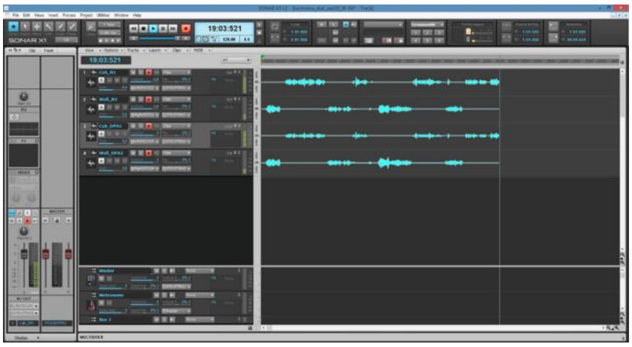

Figure 5: Screenshot of the Sonar X1 LE recording software. 1st and 3rd sound tracks – speaker A, 2nd and 4th sound tracks – speaker B.

3. Annotation specifications of the dialogue corpus

The annotation of the dialogues was carried out on 7 tiers in Praat (Boersma & Weenink 2001):

1. ort_A – orthographic and prosodic annotation, speaker A
2. DA_A – dialogue acts, speaker A
3. info_A – metadata: information about speaker, e.g. excited, information about relation between speakers, e.g. dominant, any additional information, speaker A
4. ort_B – orthographic and prosodic annotation, speaker B
5. DA_B – dialogue acts, speaker B
6. info_B – metadata, speaker B
7. agree – parts of dialogues where both speakers agree or not, information about convergence in dialogue.

The annotation tiers are described in details in Demenko and Bachan (2017). The annotation process is still in progress and so far only a few recordings sessions were fully annotated according to the abovementioned specifications.

4. Discussion and future work

In the present paper, the creation of a dialogue corpus for phonetic convergence analysis and modelling was presented. The dialogue scenarios and controlled speech prompts were shown in details and the recording method and equipment set-up in a professional studio was presented. Finally, the annotation specifications of spontaneous speech were introduced. This scenario combination and annotation specifications are novel, and promise to provide an empirical foundation for both linguistic and computational dialogue modelling of both face-to-face and man-machine dialogue by providing systematic quantitative data on convergence in a set of plausible scenarios.

5. Acknowledgements

The present study was supported by the Polish National Science Centre, project no.: 2014/14/M/HS2/00631, “Automatic analysis of phonetic convergence in speech technology systems” and was conducted in cooperation with the project partner Prof. Bernd Möbius who is the leader of a project “Phonetic convergence in Human-Machine Communication.”

References

Bachan, J. (2011a). *Modelling semantic alignment in emergency dialogue.* In: Proceedings of 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. 25-27 November 2011, Poznań, Poland, pp. 324-328.

Bachan, J. (2011b). *Communicative alignment of synthetic speech. Ph.D. Thesis.* Institute of Linguistics, Adam Mickiewicz University in Poznań, Poland.

Baumann, S. and Grice, M. (2006). *The intonation of accessibility.* Journal of Pragmatics, 38, pp. 1636-1657.

Boersma, P. and D. Weenink. (2001). *PRAAT, a system for doing phonetics by computer.* In: Glot International 5(9/10), pp. 341-345.

Carlson, R., Edlund, J., Heldner, M., Hjalmarsson, A., House, D. and Skantze, G. (2006). *Towards human-like behaviour in spoken dialog systems.* In: Proceedings of Swedish Language Techno-logy Conference (SLTC). Gothenburg, Sweden.

Edlund, J., Gustafson, J., Heldnera, M. and Hjalmarssona, A. (2008). *Towards human-like spoken dialogue systems.* Speech Communication 50(8-9), pp. 630-645.

Edlund, J., Heldner, M. and Gustafson, J. (2006). *Two faces of spoken dialogue systems.* In: Inter-speech 2006. Pittsburgh, PA, USA.

Demenko, G. and J. Bachan. (2017). *Annotation specifications of a dialogue corpus for modelling phonetic convergence in technical systems.* In: Studientexte zur Sprachkommunikation - Proceedings of 28th Conference on Electronic Speech Signal Processing (ESSV), 15-17 March 2017, Saarbrücken, Germany.

Gibbon, D., R. Moore and R. Winski. (1997). *Handbook of Standards and Resources for Spoken Language Systems.* Berlin: Mouton de Gruyter.

Giles, H. (1973). Accent mobility: *A model and some data.* Anthropological Linguistics 15, pp. 87– 105.

Giles, H., N. Coupland and J. Coupland. (1991). *Accommodation Theory: Communication, context, and consequence.* In: Giles, H., J. Coupland, and N. Coupland (Eds.): Contexts of Accommodation: Develop-ments in Applied Sociolinguistics, pp. 1 – 68, Cambridge University Press.

Gorisch, J., Wells, B. and Brown, G. (2012). *Pitch Contour Matching and Interactional Alignment across Turns: An Acoustic Investigation.* Language and Speech, 55, pp. 57-76.

Lelong, A. and G. Bailly. (2011). *Study of the phenomenon of phonetic convergence thanks to speech dominoes.* In: A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud and A. Nijholt. Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issue, Springer Verlag, pp. 280-293, LNCS.

Madonna on the cross. (photo). <http://pu.i.wp.pl/k,NjY0OTU2MzIsNDYyNzkwODE=,f,madonna_krzyz_020209.jpg>, accessed on 2016-11-30.

Oertel, C., Gustafson, J., and Black, A. (2016). *On Data Driven Parametric Backchannel Synthesis for Expressing Attentiveness in Conversational Agents.* In: Proceedings of Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction (MA3HMI), satellite workshop of ICMI 2016.

Pardo, J. S. (2006). *On phonetic convergence during conversational interaction.* Journal of the Acoustical Society of America 119, pp. 2382 – 2393.

Pickering, M. J. and S. Garrod. (2004). *Toward a mechanistic psychology of dialogue.* Behavioral and Brain Sciences 27, pp. 169 – 225.

Porzel, R. and Baudis, M. (2004). *The Tao of CHI: Towards effective human-computer interaction.* In: HLT-NAACL 2004: Main Proceedings (Boston, Massachusetts, USA, May 2 - May 7 2004), D. M. Susan Dumais and S. Roukos, Eds., Association for Computational Linguistics, pp. 209-216.

Porzel, R., Scheffler, A. and Malaka, R. (2006). *How entrainment increases dialogical efficiency.* In: Proceedings of Workshop on Effective Multimodal Dialogue Interfaces, Sydney.

Savino M., Lapertosa L. Caffò, A. and Refice, M. (2016). *Measuring prosodic entrainment in Italian collaborative game-based dialogues.* Proceedings of the 18th International Conference on Speech & Computer, Budapest 23-27 August 2016, p.476-483, LNCS Series n.9811, pp.476 – 483, Springer Verlag.

Sonar X1 LE. <https://www.roland.fi/products/sonar_x1_le/>, accessed on 2017-03-11.

van Engen, K. J., M. Baese-Berk, R. E. Baker, A. Choi, M. Kim and A. R. Bradlow. (2010). *The Wildcat Corpus of Native-and Foreign-accented English: Communicative Efficiency across Conversational Dyads with Varying Language Alignment Profiles.* Language and Speech, Vol. 53, 4, pp. 510 – 540.

Ward, A. & Litman, D. (2007). *Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora.* In: Proceedings of the SLaTE Workshop on Speech and Language Technology in Education.