Development of a Large Spontaneous Speech Database of Agglutinative Hungarian Language

Tilda Neuberger, Dorottya Gyarmathy, Tekla Etelka Gráczi, Viktória Horváth, Mária Gósy, and András Beke

Research Institute for Linguistics of the Hungarian Academy of Sciences
Departement of Phonetics, Benczúr 33, 1068 Budapest, Hungary
{neuberger.tilda,gyarmathy.dorottya,graczi.tekla,horvath.viktoria,
gosy.maria,beke.andras}@nytud.mta.hu

Abstract. In this paper, a large Hungarian spoken language database is introduced. This phonetically-based multi-purpose database contains various types of spontaneous and read speech from 333 monolingual speakers (about 50 minutes of speech sample per speaker). This study presents the background and motivation of the development of the BEA Hungarian database, describes its protocol and the transcription procedure, and also presents existing and proposed research using this database. Due to its recording protocol and the transcription it provides a challenging material for various comparisons of segmental structures of speech also across languages.

Keywords: database, spontaneous speech, multi-level annotation

1 Introduction

Nowadays the application of corpus-based and statistical approaches in various fields of speech research is a challenging task. Linguistic analyses have become increasingly data-driven, creating a need for reliable and large spoken language databases. In our study, we aim to introduce the Hungarian database named BEA that provides a useful material for various segmental-level comparisons of speech also across languages. Hungarian, unlike English and other Germanic languages, is an agglutinating language with diverse inflectional characteristics and a very rich morphology. This language is characterized by a relatively free word order. There are a few spoken language databases for highly agglutinating languages, for example Turkish [1], Finnish [2]. Language modeling of agglutinating languages needs to be different than modeling of languages like English [3]. There are corpora of various sizes, different numbers of speakers and diverse levels of transcription. TIMIT Acoustic-Phonetic Continuous Speech Corpus was created for training speaker-independent speech recognizers. This database consists of sentence reading from 630 American English speakers; includes time-aligned orthographic, phonetic and word transcriptions [4]. The Verbmobil database (of 885 speakers) was developed also in the 90's with speech technological purposes [5]. The spoken part of the British National Corpus (100 million words) [6] consists of informal dialogues that were collected in different contexts, ranging from formal business or government meetings to radio shows. The London–Lund Corpus contains 100 texts of spoken British

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 424–431, 2014.

English. The basic prosodic features, simultaneous talk, contextual comment (laughs, coughs, telephone rings, etc.) were marked in the annotation [7]. The Switchboard corpus [8] includes 2,400 telephone dialogues of 543 American English speakers. It was developed mostly for the applications in speaker identification and speech recognition. There are also some corpora of audio and transcripts of conversational speech, such as HCRC map task corpus [9] or Buckeye corpus [10], and natural meetings, such as ICSI (International Computer Science Institute) Meeting Corpus [11] or AMI (Augmented Multi-party Interaction) Meeting Corpus [12]. Although the earliest databases had consisted of written and spoken English texts, new corpora were developed also in other languages in the past decades (e.g, the German Kiel Corpus [13], Danish spoken corpus [14]. The CSJ (Corpus of Spontaneous Japanese) is one of the largest databases; it contains 661 hours of speech by 1,395 speakers including 7.2 million words [15]. EUROM1 [16] and BABEL [17] are multilingual databases, containing samples of various languages giving possibility to compare the phonetic structures of these languages using similar materials and recording protocols in all languages. Recordings of spoken Hungarian were first compiled at the beginning of the twentieth century; unfortunately, this material was destroyed. Various types of dialectical speech materials were recorded in the 1940s; these recordings were archived in the late nineties and are available for studying at the Research Institute for Linguistics of the Hungarian Academy of Sciences, RIL. The Budapest Sociolinguistic Interview contains tape recorded interviews with 250 speakers (2–3 hours each) made in the late eighties [18]. The Hungarian telephone speech database (MTBA) is a speech corpus containing read speech recorded via phone by 500 subjects. It was designed to support research and developments in the fields of speech technology [19]. The HuComTech Multimodal Database contains audio-visual recordings (about 60 hours) of 121 young adult speakers that represent North-East Hungary [20]. The developing of the largest Hungarian spontaneous speech database, BEA (the abbreviation stands for the letters of the original name of the database: BEszélt nyelvi Adatbázis ‘Speech Database ‘Speech Database’) started at the Phonetics Department of RIL in 2007. This database involves a great number of speakers who speak relatively long, contains various styles of speech materials, and has various levels of transcriptions.

2 Database Specification

At the moment of writing this paper, the total recorded material of BEA comprises 333 recordings, meaning 300 hours of speech material (approximately 4,500,000 words). The shortest recording lasts 24 minutes and 27 seconds, the duration of the longest is 2 hours, 24 minutes and 47 seconds; the average length is 51 minutes (SD: 15.8). The majority of them appear between 40 and 60 minutes. Speech materials from 184 female and 149 male speakers are available at the moment. For each recording, the following data are documented: the participant’s age, schooling, job, height, weight, whether s/he is a smoker. The youngest participant is 19 years old, while the oldest one is 90 years old. The mean age of speakers is 39 years (SD: 18.8). The majority of the participants are in their twenties and thirties.

The database contains various types of speech materials: mainly spontaneous speech, but it also includes sentence repetitions and read texts; which provides an opportunity for comparison among speech styles. The protocol consists of six modules: 1. sentence repetition (25 phonetically-rich sentences), 2. spontaneous narrative about the subject's life, family, job, and hobbies, 3. opinion about a topic of current interest, 4. directed spontaneous speech; summary of content of two heard text, 5. three-party conversation, and 5. reading of sentences and text (for further details see [21]). In 95% of all recordings, the interviewer was the same young woman. Recordings are invariably made in the same room, under identical technical conditions: in the sound-proof booth of the Phonetics Department, specially designed for the purpose. The size of the room is 3.4 × 2.1 × 3.0 m. The walls of the room are provided with a sound-absorbing layer in order to avoid reverberation. The degree of sound damping as compared to the outside environment is 35 dB at 50 Hz, and ≥ 65 dB above 250 Hz. The recording microphone is AT4040. Recording is made digitally, direct to the computer, with GoldWave sound editing software, with sampling at 44.1 kHz (storage: 16 bits, 86 kbytes/s, mono).

3 Transcription, Segmentation and Labeling

The BEA database has three types of transcription. The first was done in MS Word, the second in Transcriber, the third is being done in Praat. This chapter introduces the first two in nutshell, as they have been introduced already in details [21], and the third one is to be described in details first time in this study.

1. The primary transcription in MS Word (.doc format) is based on the orthography but without punctuation. The participants are uniformly abbreviated in these transcriptions as A (subject), T1 (interviewer and first conversational partner), T2 (second conversational partner). The proper names are capitalized, and some phenomena are marked: disfluencies (bold), hesitations, hummings, and other non-verbal noises like laughter (exclamation mark), as well as speaking simultaneously (parentheses), perceived pauses (❑) (see Fig. 1). 47% of the recordings were transcribed in this format.

T2 ([1]![2] **oo**[3]r dono[4] [I don't know][5])	→ [1]simultaneous speech, [2]breath, [3]lengthening, [4]causal word form [5]intended form/expression
A (**soo**)	
A long ago the old people **eer**[6]![7] so! said that! th**aat**!	→ [6]hesitation, [7]breath/noise...
probem[8] [problem] tha [that] **tho**[9] [though] any degrees you should learn a manual occupation	→ [8]slip-of-the-tongue, [9]unfinished word
A (my son!)	
T1 (yes yes)	
T2 (mhm[10])	→ [10]humming
A and this this vie**ww eer** in the past forty years	
A (disappeared)	
T2 (did disappear)	

Fig. 1. Sample fragment of conversation in primary transcription (translated into English, with transcription marks)

2. Transcription made in Transcriber (`http://trans.sourceforge.net`) is basically time-aligned pause-to-pause labelling, also follows the rules of Hungarian orthography but without punctuation. The label boundaries are set at approximately the middle of at least 200 ms long pauses. Even longer pauses are marked separately. Noises from the speakers or the environment, non-speech events (like laughter, cough), hummings, hesitations, unfinished words, simultaneous utterances, unintelligible speech, and other phenomena are marked with special abbreviations and codes. The speakers are identified using the same characters as in the Word-transcriptions. 51% of the recordings were transcribed using this software.

3. The third type of transcription is done in Praat (`http://praat.org`) at several levels (Fig. 2). This transcription is being done at present. Criteria and rules were developed in the second half of 2013, and the transcription procedure started by trained annotators. The transcription includes 9 levels, where the first three levels include the interviewer's speech, the second three levels include the subject's talk, while the last three levels (only in part 5: conversation) are devoted to what the second conversational partner said. The first level of each speaker includes pause-to-pause labels in that speech samples are transcribed in orthography without punctuation. The second level of each speaker means word-level segmentation. The third levels are speech sound labels. Some specific simplifications are defined. Neither hyphens, nor capitals are used as opposed to the orthographical requirements since both of them have special functions. Silent ('SIL') and filled (e.g., 'M') pauses are marked in separate labels in each row of the speaker whose speech sample they belong to. When the speaker is not speaking but listening to the other(s), their lines are marked by 'PAUSE'. Unintelligible or noisy speech segments are labelled in each row as unusable parts. Simultaneous speech samples are not transcribed, but marked in each row just as overlapping speech ('E'). These transcriptions also include non-speech events (e.g. humming, laughter, sigh), disfluency phenomena, speech errors, word fragments. Slip-of-the-tongue phenomena are written as pronounced at the pause-to-pause level, and the intended word is added in square brackets. The word level includes the intended word and the sound level includes the pronounced speech sounds.

In the sound-level annotation the segment label set is phonetic and has several rules that concern specific problematic realizations. Here we give some examples. In cases where a silent pause is followed by a voiceless closure phase (of p, t, c, k, ts, tS), the boundary of the pause is consensually placed 30 ms before the first closure release. The cases, where a vowel is followed by a consonant at the phonological level that does not appear in the pronunciation but influences the realization of the vowel, are marked in the vowel label. Irregular and breathy voice and aspiration are marked also in the sound label ('Y', 'W', 'H', respectively). Further special cases are also appropriately discussed in the instructions for the transcribers.

4 Research Based on BEA

In this chapter the usefulness of BEA will be evaluated from a point of view of speech science. Research has been initiated in the following areas of phonetics, psycholinguistics and speech technology: the segmental structure of speech, coarticulation, supraseg-

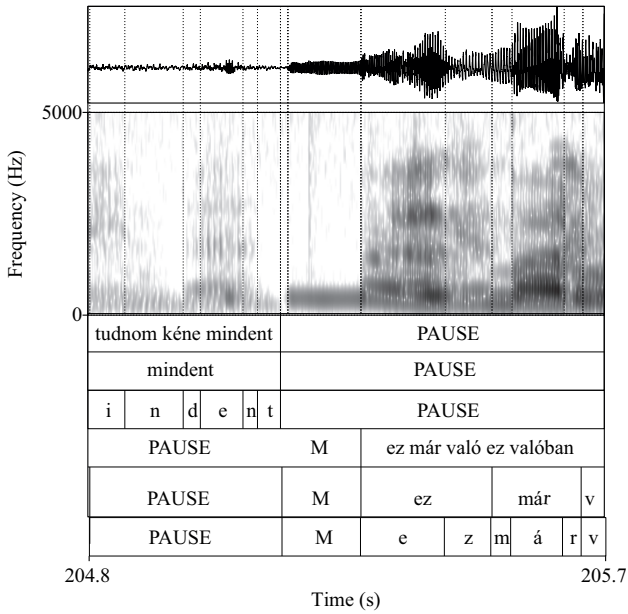

Fig. 2. A sample of turn taking in Praat annotations (M = hesitation).

mental features of speech, fluency of speech, temporal factors, disfluencies, non-verbal vocalizations, automatic classification of various speech phenomena, speech detection, overlapping speech detection, speaker diarization.

A collection of studies has been published in Hungarian focusing on recent investigations where this database was used [22]. The durations and formant frequencies of the Hungarian vowels (more than 10,000 tokens) were measured in spontaneous speech of BEA [23,**?**,**?**]. One of the questions of the various investigations was how different vowels can be discriminated in spite of the large overlaps of the formant frequencies. The results showed that the accuracy of J48 classifier was higher depending on the horizontal tongue movements (87%) than depending on the vertical ones (69.7%). Fricative phoneme realizations (total duration of the consonants, duration of the voiced part, mean HNR, COG and other features) and the frequency of neutralization or weakening of the voicing oppositions were analyzed using both spontaneous and read speech samples of eight speakers of BEA [25]. Multilayer Perceptron neural network method was used for automatic classification. In a recent study [26] an attempt was made to define various units of spontaneous narratives and capture objective acoustic-phonetic properties of boundary marking. The results showed that (i) the majority of the speakers organize their narratives in similar temporal structures, (ii) thematic units can be identified in terms of certain prosodic criteria, and (iii) there are statistically valid correlations between factors like the duration of phrases, the word count of phrases, the rate of articulation of phrases, and pausing characteristics. Several investigations focused on the examination of speech planning and self-monitoring mechanisms by analyzing disfluency phenomena. An analysis [27] of the frequency and phonetic characteristics

of anticipations and perseverations (in spontaneous speech samples by twenty-seven speakers) revealed that higher-organized units could drift away from their planned position to a relatively longer distance in time than lower-organized units while the latter tended to do so more frequently than the former. Temporal patterns confirmed that the speech production mechanism controls pre-planning more successfully. False starts and false words were investigated in a large amount of spontaneous speech samples (16-hour speech material consists of narratives of 70 adults) [28]. The results confirmed that false starts occur more frequently than false words, which indicates the appropriate functioning of the covert self-monitoring. The duration of the editing phase is affected by its structure the same way in both types of the analyzed disfluencies; and depends on the type of the word, and on the relation between the reparandum and the repair. A PhD thesis addressed the topic of speaker diarization for 100 spontaneous conversations of BEA [29]. The presented speaker diarization system was based on unsupervised learning method which could be easily adapted to another speech corpus. The best result (DER: 28.71%) was yielded using BIC-base method where the penalty value was 1, the features were MFCC(2,5–3,5) and the system contained the VAD and overlapping detection algorithm as well. Spontaneous conversations (also in the BEA database) frequently contain various non-verbal vocalizations such as laughter. The sound sequence of laughter may acoustically resemble to speech sounds; F0, formant structure, and RMS amplitude of laughter seem to be rather speech-like. There was an attempt to develop an accurate and efficient method to differentiate laughter from other speech events [30]. The results showed that the GMM-SVM system trained on acoustic parameters, MFCC and PLP could be a particularly good method for solving this problem.

5 Conclusion and Future Work

There are many research possibilities provided by BEA. It records the contemporary state of spoken Hungarian, providing the foundation for later comparative studies of linguistic change. In a number of areas like phonetics, laboratory phonology, speech technology, psycholinguistics, applied speech research, pragmatics, spontaneous speech grammatics, socio-phonetics, speaker identification (forensic phonetics) or speech-based medical diagnostics, most examinations can only be done on the basis of a large amount of speech material meeting the criteria of database technology. Although the BEA corpus was created to study phonetic aspects of speech, it should be useful to scientists interested in many other (linguistic) aspects of spontaneous speech.

The database is available for any researcher by contacting the developers. The files are not uploaded to the internet in order to warrant the speakers' privacy rights. However, we are planning the elaboration of an open access infrastructure, which provides an access to the corpus (both recordings and annotation) with privacy and security conditions. The corpus will be made available to the scientific community when transcription is completed.

Acknowledgments This work was supported by OTKA 108762.

References

1. Mengusoglu, E., Deroo, O.: Turkish LVCSR: Database preparation and language modeling for an agglutinative language. In: IEEE International Conference On Acoustics Speech And Signal Processing, Vol. 6, IEEE, 1999, pp. 4018–4018. (2001)
2. Seppänen, T., Toivanen, J., Väyrynen, E.: MediaTeam speech corpus: a first large Finnish emotional speech database. In: Proceedings of the Proceedings of XV International Conference of Phonetic Science, pp. 2469–2472. (2003)
3. Mihajlik, P., Fegy´o, T., T¨uske, Z. and Ircing, P.: A morphographemic approach for the recognition of spontaneous speech in agglutinative languages - like Hungarian. In: Proc. Interspeech 2007, Antwerp, Belgium, pp. 1497–1500, (2007)
4. Keating, P., Byrd D., Flemming, E., Todaka Y.: Phonetic analyses of word and segment variation using the TIMIT corpus of American english. In: Speech Communication 14 (2).pp. 131–142. (1994)
5. Bael, C. V., Boves, L., van den Heuvel, D., Strik, H.: Automatic phonetic transcription of large speech corpora. In: Journal of Computer Speech and Language 21 (4). pp. 652–668. (2007)
6. Aston, G., Burnard L.: The BNC Handbook. Exploring the British National Corpus with SARA. Oxford University Press. (1998)
7. Svartvik, J., ed.: The London Corpus of Spoken English: Description and Research. Lund Studies in English, 82. Lund: Lund University Press. (1990)
8. Godfrey, J. J., Holliman, E. C., Daniel J.: SWITCHBOARD: telephone speech corpus for research and development. In: Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., Vol. 1. pp. 517–520. (1992)
9. Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S.,…Weinert, R.: The HCRC map task corpus. In: Language and speech, 34(4), pp. 351–366. (1991)
10. Pitt, M. A., Johnson, K. Hume, E., Kiesling, S., Raymond, W.: The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. In: Speech Communication 45. pp. 89–95. (2005)
11. Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., ... Wooters, C.: The ICSI meeting corpus. In: Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on Vol. 1, pp. 364–367. (2003)
12. Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., ... Wellner, P.: The AMI meeting corpus: A pre-announcement. In: Machine learning for multimodal interaction, Springer Berlin Heidelberg. pp. 28–39. (2006)
13. Kohler, K. J., Pätzold, M., Simpson, A. P.: From the acoustic data collection to a labelled speech data bank of spoken Standard German. In: Arbeitsberichte des Instituts fär Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK), 32, pp. 1–29. (1997)
14. Grønnum, N. A Danish phonetically annotated spontaneous speech corpus (DanPASS). In: Speech Communication, 51 (7), pp. 594–603. (2009)
15. Maekawa, K.: Corpus of Spontaneous Japanese: Its design and evaluation. In ISCA IEEE Workshop on Spontaneous Speech Processing and Recognition (2003).
16. Chan, D., et al.: EUROM: a spoken language resource for the EU. In: Proceedings of the 4th European Conference on Speech Communication and Speech Tecnology,Eurospeech '95, (Madrid) 1, pp. 867–880. (1995)
17. Roach, P., Arnfield, S., Barry, W. J., Baltova, J., Boldea, M., Fourcin, A., ... Vicsi, K.: BABEL: an eastern european multi-language database. In: ICSLP. (1996)
18. Váradi, T.: A Budapesti Szociolingvisztikai Interjú. In: Kiefer F, Siptár P. (ed.). A magyar nyelv kézikönyve Akadémiai Kiadó, Budapest, pp. 339–359. (2003)

19. Vicsi, K., Tóth, L., Kocsor,. Gordos, G., Csirik, J.: MTBA – magyar nyelvű telefonbeszéd-adatbázis. Híradástechnika 8. pp. 35–39. (2002)
20. Pápay, K.: Designing a Hungarian multimodal database – speech recording and annotation. In: Anna Esposito et al.: Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues. Berlin Heidelberg: Springer-Verlag, pp. 403–11. (2011)
21. Gósy, M.: BEA A multifunctional Hungarian spoken language database. In: The Phonetician, 105(106), pp. 50–61. (2012)
22. Gósy M. (ed.): Beszéd, adatbázis, kutatások. Akadémiai Kiadó, Budapest. (2012)
23. Gráczi, T. E., Horváth, V.: A magánhangzók realizációja spontán beszédben. Beszédkutatás 2010. pp. 5–16. (2010)
24. Beke, A., Gósy, M.: Characteristic and spectral features used in automatic prediction of vowel duration in spontaneous speech. In: Institute of Electrical Electronics Engineers (eds.): CogInfoCom 2012: 3rd International Conference on Cognitive Infocommunications. pp. 65–71. (2012)
25. Gráczi, T. E. – Beke, A.: Fricatives in spontaneous speech. In: ExAPP 2013, Copenhagen, March, 20–22. (2013)
26. Beke, A., Gósy, M., Horváth,V.: Temporal variability in spontaneous Hungarian speech. In: Proceedings of 6th Language Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. Poznan, December, 7–9. pp. 219–223. (2013)
27. Gósy, M., Gyarmathy, D., Horváth, V.: Improper activation and monitoring failures in speech planning. In: Govor / Speech 29/1. pp. 3–22. (2012)
28. Gyarmathy, D., Neuberger, T.: Self-monitoring strategies: the factor of age. In: Presentation at the 19th International Congress of Linguists, Geneva, July, 21–27. (2012)
29. Beke A.: Automatic speaker diarization in Hungarian spontaneous conversations. PhD thesis. ELTE, Budapest. (2013)
30. Neuberger, T., Beke, A.: Automatic laughter detection in spontaneous speech using GMM-SVM method. In: Habernal, I., Matousek, V. (eds.): Text, Speech and Dialogue. 16th International Conference, TSD 2012, Pilsen, Czech Republic. Springer. pp. 113–120. (2013)