A Corpus of Spontaneous Multi-Party Conversation in Bosnian Serbo-Croatian and British English

Emina Kurtic1,2, Bill Wells2, Guy J. Brown1, Timothy Kempton1, Ahmet Aker1
Department of Human Communication Sciences1 and Department of Computer Science2, University of Sheffield, UK
email: e.kurtic@sheffield.ac.uk, b.wells@sheffield.ac.uk, g.brown@dcs.shef.ac.uk, t.kempton@dcs.shef.ac.uk, a.aker@dcs.shef.ac.uk

Motivation
- Are the temporal and phonetic characteristics of overlapping talk similar across languages?
- To answer this question, need high quality recordings made under the same conditions in two languages.

Characteristics of the Corpus
- Collected in British English (BE) and Bosnian Serbo Croatian (BSC) to allow comparative investigations;
- High quality audio and video recordings;
- Naturally occurring, face-to-face talk;
- Non-institutional, since spoken exchange in institutional meetings can be influenced by the agenda etc. [5];
- Each participant recorded on a separate audio channel to allow reliable analysis of acoustic features.

Data Collection and Recording
- Participants were native speakers of each language, three female and one male.
- Existing friendship groups (students).
- Informal setting in univeristy room (e.g., participants could eat and drink).
- 3 hours of recordings for each language.
- Digital audio recordings made on MacBook Pro, using AudioDesk software and MOTU 8Pre Firewire audio interface.
- Participants recorded with Sennheiser ME 3-N cardioid headsets.
- Omnidirectional recording with PZM microphone. BE meetings also recorded with an array of 8 microphones [4].
- Video recordings made with Canon MN600 camera. Additional Canon XM2 camera used for the BE recordings.

Segmentation
- Speech was segmented into turn constructional units (TCUs) [5].
- TCU is a minimal constituent of a speaker turn; syntactically and pragmatically complete, building block used by talkers.

Transcription
- Transcripts consist of four tiers:
  - Comments
  - Orthography
  - Non-speech sounds
  - Uncertain

- Segmentation at the TCU, word, and phone levels.
- Forced alignment used to generate segmentation at word and phone levels.
- For BE, forced alignment used acoustic models and pronunciation dictionary trained on the AMI corpus [1].
- The 20ms error for BE (proportion of boundaries placed more than 20ms from ground truth boundary) was 35%.
- No resources available for BSC, so used cross-language forced alignment [2].
- Alignment used recognisers trained on Czech, Russian, Hungarian or American English (AE): BSC phones automatically mapped to the closest phone of the target language.

- Best alignment with Russian recogniser, significantly better than Hungarian and AE (Wilcoxon signed rank test, p<0.01).
- F2 and F3 talk about the practice that family attends the public defence of theses (viva).
- F2 doesn’t want anyone to come to the viva, F3 argues that it’s important for the family.
- They overlap each other many times in the extract.

Corpus Availability
- The corpus will be released from July 2012 under the Creative Commons Licence.
- Accessible via a web-based search engine which allows download of audio and video segments.
- Search for participants, annotated items, regions of overlapping speech etc.

Conclusions and Future Work
- Novel audio-visual corpus of spontaneous multi-party conversations in BSC and BE.
- Cross-language techniques are viable for segmentation of BSC, for which few resources exist.
- Overlapping talk is common in the corpus.
- Current work is investigating how prosodic resources are used to signal competitive vs. noncompetitive overlaps in the two languages [3].

Acknowledgments

Supported by AHRC. Thanks to Matt Gibson and Thomas Hain for assistance with forced alignment.

References

Reading setup for Bosnian Serbo Croatian at the University of Tuzla, Bosnia and Herzegovina (top) and setup for British English recordings at the University of Sheffield UK (bottom).