Developing an Annotation System for Communicative Functions for a Cross-Layer ASR System

Barbara Schuppler
SPSC Laboratory
Graz University of Technology, Austria
b.schuppler@tugraz.at

Anneliese Kelterer
Department of Linguistics,
University of Graz, Austria
anneliese.kelterer@uni-graz.at

Abstract

The investigation of conversational speech requires the close collaboration of linguists and speech technologists to develop new modeling techniques that allow the incorporation of various knowledge sources. This paper presents a progress report on the ongoing interdisciplinary project ”Cross-layer language models for conversational speech” with a focus on the development of an annotation system for communicative functions. We discuss the requirements of such a system for the application in ASR as well as for the use in phonetic studies of talk-in-interaction, and illustrate emerging issues with the example of turn management.

1 Cross-layer language models for conversational speech

In the last decade, conversational speech has received a lot of attention among speech scientists. Accurate automatic speech recognition (ASR) systems are essential for conversational dialogue systems, as these become more interactional and social rather than solely transactional (Baumann et al., 2016). Linguists study natural conversations, as they reveal additional insights to controlled experiments with respect to how speech processing works. Investigating conversational speech, however, does not only require the application of existing methods to new data, but also the development of new categories and modeling techniques, and the inclusion of new knowledge sources.

Here, we present an ongoing interdisciplinary project with two main aims: (1) The project aims at increasing our understanding of how phonetic (and especially prosodic) variation is related to the semantic context and to communicative functions in conversations. For this purpose, we will conduct phonetic corpus studies and perception experiments, both based on data drawn from conversational speech corpora.

(2) Whereas traditional language models (LMs) are trained on text only, we aim at incorporating information on the phonetic variation of words in LMs and at relating this information to the semantic context and to the communicative functions in conversation. This approach to LMs is in line with the theoretical model proposed by Hawkins and Smith (2001), in which the perceptual system accesses meaning from speech by using the most salient sensory information from any combination of levels/layers of formal linguistic analysis. Such a model is reminiscent of the cross-layered optimization principle in wireless communications (Shakkottai et al., 2003). It was introduced as an alternative to the Open Systems Interconnection (OSI) model, where one layer provides services only to its upper layer while exclusively receiving services from the layer below. With the term *cross-layer*, we refer to our view of how humans access meaning and to the system architecture of the envisioned ASR system.

Figure 1 shows the architecture of the ASR system which is currently being developed. Boxes in white show components that have already been developed (Schuppler et al., 2017; Linke et al., 2020; Schuppler and Ludusan, 2020). Those in gray are currently being developed. The LM proposed is aware of the communicative history and dynamics of the conversation (in Figure 1 referred to as ’cache’). Our current ASR experiments show that WERs heavily depend on pronunciation variation, articulation rate, overlapping speech and semantic and syntactic complexity, which in turn strongly correlate with communicative functions. Our knowledge-based approach to LMs is contrary to recent work on end-to-end ASR systems (e.g., Ito et al., 2017), because in addition to improving ASR, we also aim at increasing our knowledge on human speech processing.

One important aspect of our work is its interdis-

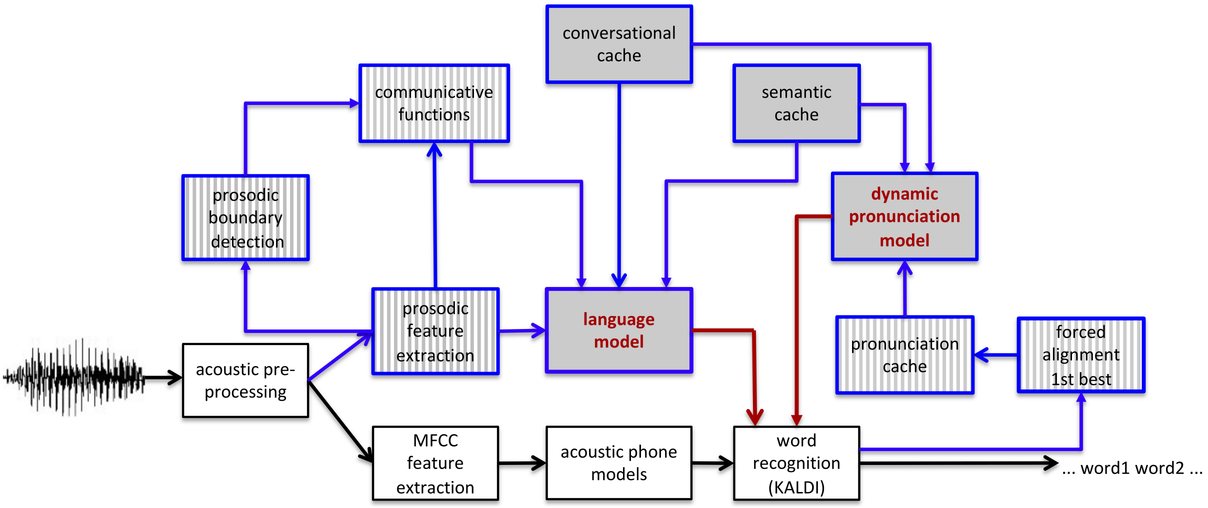

Figure 1: Architecture for an ASR system using a communicative-functions aware language model.

ciplinarity work flow. We create cross-layer LMs which will be tested in ASR systems. In doing so, we will not only investigate which contextual, lexical and acoustic cues work well for speech recognition, but we will also interpret them phonetically. Subsequently, corpus and perception studies will be designed to investigate which of the cues used by the ASR system are also relevant for human speech perception, and which additional cues used by humans might increase ASR performance (e.g., Schuppler et al., 2010). The phonetic studies will be facilitated by ASR technology, i.e., we use tools for the annotation of data, for acoustic feature extraction and we apply advanced statistical methods. Gained phonetic knowledge will again be incorporated into the ASR system. For this interdisciplinary workflow, it is thus necessary to develop an annotation system of communicative functions which is suitable for both phonetic studies and the incorporation into an ASR system.

2 GRASS Corpus

The Graz corpus of Read and Spontaneous Speech (GRASS) contains recordings of spontaneous dialogues of one hour each. They were recorded with 19 pairs of native speakers of (eastern) Austrian German who were friends, couples or family members, resulting in a casual speaking style. The orthographic transcriptions include annotations of disfluencies, breathings and laughter (Schuppler et al., 2014, 2017). Parts of the corpus have been segmented on word and phone-level and were manually annotated prosodically following the KIM system (IPDS, 1997). We have built tools for the detection of prosodic boundaries (Schuppler and Ludusan, 2020) and for the classification of prominence levels (Linke et al., 2020). These tools were created such that they can (1) facilitate the annotation of the not yet annotated parts of GRASS in a semi-automatic procedure, and (2) can be incorporated into the ASR system shown in Figure 1. For the communicative-functions layer of annotations, we also aim to build a tool that serves both mentioned purposes.

3 Annotation of Communicative Functions

For GRASS, we need an annotation scheme that is suitable for speech in naturally occurring conversations. Thus, we take notions from Conversation Analysis (CA), a discipline that focuses on speaker behaviour (rather than, e.g., intentions or intuitions) and stresses the importance of the sequential context for the analysis of speech. Most annotations of communicative functions in the literature are restricted to a limited set of data tailored to a specific investigation (e.g., Ward, 2004; Gravano et al., 2007). One exception are the dialog act categories used to annotate the Switchboard Corpus (Jurafsky et al., 1998; Calhoun et al., 2010). Other corpora that are transcribed in CA terms are searchable for words/lemmata, but not annotated for communicative functions (e.g., *DGD*; Schmidt, 2014). We aim at creating annotations of communicative functions for whole conversations in GRASS. The communicative functions annotations will be used for (1) improving ASR with knowledge about turn-taking, feedback particles with different functions and speaker alignment (e.g., agreement and disagreement), and how they relate to prosody and pronunciation variation; and (2) studying the function-phonetics mapping for various questions in the tradition of Phonetics of Talk-in-Interaction (PTI; Ogden, 2012). Given these two applications, our annotation system has to meet the requirements of (1) annotation consistency, (2) PTI perspective, and (3) ASR application.

Annotation consistency In comparison to PTI studies (e.g., Gorisch et al., 2012; Sikveland, 2012; Zellers, 2016), in which annotations are performed mainly by one or two experts, in our project, large amounts of data are being annotated by a team of approx. 2-4 student assistants. To obtain a high annotation quality and consistency, it is important to keep the annotation task as simple as possible. A way to achieve this is by splitting the annotation into various levels, each of a lower complexity. Another motivation for simplifying labelling tasks for human annotators is that the consistent segmentation and labelling of units are essential to ensure good automatic detection of categories.

PTI perspective For the investigation of prosodic and segmental phonetic variation in an integrated approach such as proposed by Zellers and Post (2011), the annotation of communicative functions has to be methodologically sound following principles of Conversation Analysis (CA). One domain we employ in our annotation scheme is potential transition relevance places (TRP) in terms of *points of potential syntactic completion* (PCOMP). While TRPs are undoubtedly also determined by prosody (e.g., Selting, 1996), it is less clear what constitutes potential phonetic completion. Therefore, even studies within PTI use only syntactic criteria to identify potential TRPs in their investigations of turn management (e.g., Zellers, 2016; Local and Walker, 2012). For the ASR system, the annotation of PCOMPs might pose problems, in particular in cases in which they do not coincide with pauses. Since these domains are not well-defined in terms of prosody, they are harder to detect. In cases in which a pause belongs to the same unit as the stretches of speech around it (e.g., when a speaker makes a pause in the middle of a sentence, cf., Figure 2), units are difficult to recognize automatically.

ASR application For ASR, we want to use communicative functions and prosody features to improve word recognition. Thus, the word level is not available for the identification of PCOMPs in the speech stream. For the application in our ASR system, it is important that boundaries and labels can be detected on the basis of spectral and prosodic features only, as communicative functions are being detected before word-level recognition is done. Moreover, the preference is towards a small number of labels, as a large number of categories (e.g. 42 dialog act categories in Jurafsky et al., 1998, 24 stance type labels in Freeman, 2019) will lead to a high level of confusion in the automatic classification process. From an ASR point of view, the annotation of *Inter-Pausal Units* (IPU) is a viable option, since they are clearly defined and easily detectable in recordings without much background noise. If the minimal pause length is defined, the only misidentification might be extremely long plosive closure durations (e.g., in hesitations). Mismatches between communicative functions and IPUs might cause problems, particularly if one IPU includes several communicative functions, or if a communicative function stretches over more than one IPU.

Annotation labels The set of annotation labels should be suitable for the description of entire conversations without encompassing too many categories in order to reduce potential confusion by the annotator or the ASR system. For turn management, we base our set of labels on four categories used in Zellers (2016), which are defined in terms of CA, i.e., according to the behaviour of participants in the conversation: *Hold* (same speaker continues talking), *Change* (speaker change), *Question* (speaker transfers the turn to another speaker), and *Hearer Response Tokens* (e.g., backchannels; cf., Sikveland (2012)). On the IPU level, three additional labels were necessary to capture incomplete structures before pauses; for incomplete turn-holds (cf., Figure 2b), turn-changes, and in turn competition when one speaker interrupts himself/herself to cede the turn to the other speaker. The annotation of intervals on the PCOMP level is more finegrained. Thus, we added six labels to the system used in Zellers (2016). We subdivided *Hold* depending on the following context (continuation of syntactic structure vs. new sentence). For the same reason as on the IPU level, we added a label for incomplete turn-changes. A label for incomplete turn-holds is not necessary because no boundaries are set until a PCOMP is reached (cf., Figure 2a). We added labels for collaborative finishes to cap-

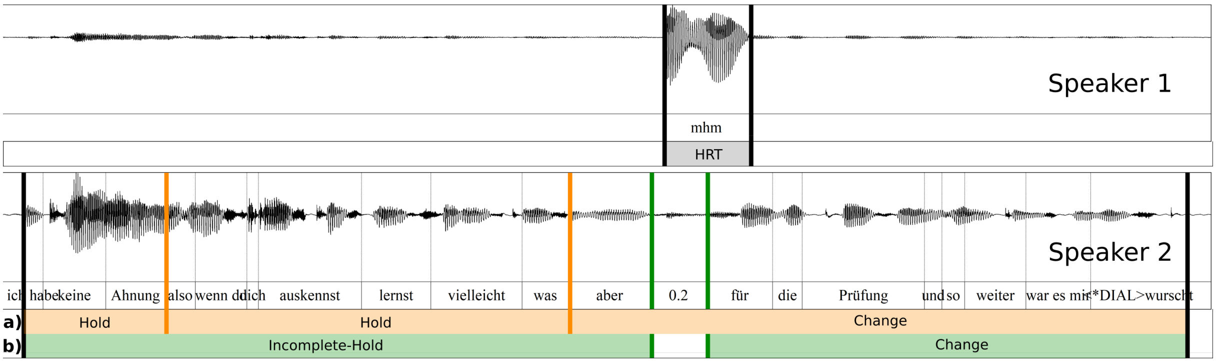

Figure 2: Time-aligned annotation of Speaker 2's turn (engl. 'I have no idea, so if you know your way around maybe you can learn something, but (0.2) for the exam and so on I didn't care.') a) at PCOMPs (orange); and b) of IPUs (green). Speaker 1 aligns his hearer response token with Speaker 2's pause after <aber>.

ture when a syntactic construction stretches over two speakers, and for discourse particles and hesitations that occur at PCOMP boundaries. Finally, we added a label for self-interruptions with subsequent rephrasing. These points in a speaker's turn are not technically PCOMPs, but they are often marked by an abrupt interruption of articulation and the syntactic reset after this point can be relevant for ASR.

Figure 2 shows an example of how PCOMP and IPU annotations are mapped onto each other. In this example, Speaker 2 holds his turn by making a pause at a point of "maximum grammatical control" (Schegloff, 1998: 241; labelled as *Incomplete-Hold* on tier b) after the introduction of a new sentence by <aber>, and completes his turn after the pause. There are two PCOMPs leading up to the pause (labelled as *Hold* on tier a), neither of which give the impression of being complete based on prosody (i.e., slightly rising pitch in <Ahnung> and 'rush-through' in <was>). Even though a pause is produced after <aber>, the next PCOMP is reached only after <wurscht>. Thus, the whole sentence starting with <aber> is grouped into one PCOMP chunk, regardless of any pauses. Speaker 1 times his backchannel (labelled as *Hearer Response Token*) with the pause rather than with the PCOMP just before <aber>. It is predominantly short hearer response tokens that are aligned with pauses at syntactically incomplete positions while participants almost never self-select to produce a longer turn in these positions.

Currently, 90 minutes in 15 conversations have been annotated at the IPU level and the last revision of these labels is in progress. On the PCOMP level, 60 minutes in 12 conversations are being annotated. These annotations are useful for the goals described above, i.e., for application in ASR and for phonetic studies, as well as for the investigation of various hypotheses about the time alignment of hearer response tokens and self-selection.

Outlook

An iterative annotation process while creating manual annotations and developing a classification tool based on acoustics will reveal more fine-grained categories (e.g., a distinction between PCOMPs that are prosodically marked as complete vs. prosodies overarching several PCOMPs). The annotation of more acoustically based categories will, in turn, improve recognition of categories. For instance, we can investigate the prosody at the end of IPUs. In a preliminary study, we performed a Random Forest classification of *Hold*, *Incomplete-Hold* and *Change* on the basis of acoustic features. An analysis of the highest ranked features in the Random Forest with linear mixed effects regression models indicated that *Incomplete-Hold*s (cf., Figure 2b) are characterized by a lower speech rate and a flatter F0 curve at the end. *Hold*s and *Change*s, on the other hand, were not consistently distinguished by prosody. The planned perception experiment of these categories will give us further insights into prosodically different kinds of turn-holds and turn-changes. The developed classifier of communicative functions will aid the annotation process by providing labels for semi-automatic annotations and will also be incorporated into our ASR system to improve word recognition.

Acknowledgments

B. Schuppler was funded by grant V-638-N33, and A. Kelterer by grant P-32700-N, both from FWF.

References

Timo Baumann, Casey Kennington, Julian Hough, and David Schlangen. 2016. Recognising conversational speech: What an incremental ASR should do for a dialogue system and how to get there. In *Proc. IWSDS*, pages 1–12.

Sasha Calhoun, Jean Carletta, Jason Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The nxt-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation Journal*, 44:387–419.

Valerie Freeman. 2019. Prosodic features of stances in conversations. *Laboratory Phonology: The Journal of the Association for Laboratory Phonology*, 10(1):1–20.

Jan Gorisch, Bill Wells, and Guy J. Brown. 2012. Pitch contour matching and interactional alignment across turns: An acoustic investigation. *Language and Speech*, 55(1):57–76.

Augustín Gravano, Stefan Benus, Julia Hirschberg, Shira Mitchell, and Ilia Vovsha. 2007. Classification of discourse functions of affirmative words in spoken dialogue. In *Interspeech*, pages 1613–1616.

Sarah Hawkins and Rachel Smith. 2001. Polysp: a polysystemic, phonetically-rich approach to speech understanding. *Italian Journal of Linguistic*, 13(1):99–188.

IPDS. 1997. CD-ROM: The Kiel Corpus of Spontaneous Speech, vol i- vol iii. Available at http://www.ipds.uni-kiel.de/forschung/kielcorpus.de.html (last viewed 25/11/2016).

Hitoshi Ito, Aiko Hagiwara, Manon Ichiki, Takeshi Mishima, Shoei Sato, and Akio Kobayashi. 2017. End-to-end speech recognition for languages with ideographic characters. In *Proc. APSIPA Annual Summit and Conference*.

Dan Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Tailor, and Carol Van Ess-Dykema. 1998. *Switchboard Discourse Language Modeling Project Report*, volume 1. Center for Speech and Language Processing, Johns Hopkins University, Baltimore.

Julian Linke, Anneliese Kelterer, Markus Dabrowsky, Dina El Zarka, and Barbara Schuppler. 2020. Towards automatic annotation of prosodic prominence levels in Austrian German. In *Proceedings of Speech Prosody 2020*, pages 1000 – 1004.

John Local and Gareth Walker. 2012. How phonetic features project more talk. *JIPA*, 42:255–280.

Richard Ogden, editor. 2012. *The Phonetics of Talk in Interaction, Special Issue in Language and Speech 55(1).*

Emmanuel A. Schegloff. 1998. Reflections on studying prosody in talk-in-interaction. *Language and Speech*, 41(3-4):235–263.

Thomas Schmidt. 2014. Gesprächskorpora und gesprächsdatenbanken am beispiel von folk und dgd. *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion*, 15:196–233.

Barbara Schuppler, Mirjam Ernestus, Wim van Dommelen, and Jacques Koreman. 2010. Predicting human perception and ASR classification of word-final [t] by its acoustic sub-segmental properties. In *Proceedings of Interspeech*, pages 2466 – 2469.

Barbara Schuppler, Martin Hagmüller, and Alexander Zahrer. 2017. A corpus of read and conversational Austrian German. *Speech Communication*, 94C:62–74.

Barbara Schuppler, Martin Hagmüller, Juan Cordovilla, and Hannes Pessentheiner. 2014. Grass: The graz corpus of read and spontaneous speech. In *9th edition of the Language Resources and Evaluation Conference*, pages 1465–1470.

Barbara Schuppler and Bogdan Ludusan. 2020. An analysis of prosodic boundary detection in German and Austrian German read speech. In *Proceedings of Speech Prosody 2020*, pages 990 – 994.

Margaret Selting. 1996. On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. *Pragmatics*, 6(3):371–388.

Sanjay Shakkottai, Theodore S. Rappaport, and Peter C. Karlsson. 2003. Cross-layer design for wireless networks. *IEEE Communications Magazine*, 41(10):74–80.

Rein Ove Sikveland. 2012. Negotiating towards a next turn: Phonetic resources for 'doing the same'. *Language and Speech*, 55(1):77–98.

Nigel Ward. 2004. Pragmatic functions of prosodic features in non-lexical utterances. In *Speech Prosody*, pages 325–328.

Margaret Zellers. 2016. Prosodic variation and segmental reduction and their roles in cuing turn transition in swedish. *Language and Speech*, 60(3):454–478.

Margaret Zellers and Brechtje Post. 2011. Combining formal and functional approaches to topic structure. *Language and Speech*, 55(1):119–139.