The Kiel Corpora of "Speech & Emotion" - A Summary

Oliver Niebuhr[1], Benno Peters[2], Rabea Landgraf[2], Gerhard Schmidt[3]

[1] *Dept. of Design & Communication, IRCA, University of Southern Denmark, E-Mail: olni@sdu.dk*
[2] *General Linguistics, Kiel University, E-Mail: p.beters@phonexis.de; landgraf@isfas.uni-kiel.de*
[3] *Digital Signal Processing and System Theory, Kiel University, E-Mail: gus@tf.uni-kiel.de*

Introduction

It is not long ago that research on speech communication was solely concerned with basic structural issues of allophonic variation, tune grammar, and boundary signals. Many issues are still far from being fully understood, even for Western European languages, but we have gained enough knowledge to start digging deeper into the social and interactional aspects of speech that actually drive communication and are coded in complex segmental and prosodic details. This shift in research focus manifests itself also in speech corpora and is undoubtedly boosted by the rapidly growing number of speech technology applications that sneak in every corner of our life. Apart from the fact that speech corpora seem to become constantly larger (for example, in order to properly train self-learning speech synthesis/recognition algorithms), the *content* of speech corpora also changes. In particular, recordings of isolated logatomes, words or sentences are successively supplanted by more realistic, interactive, and informal speech-production tasks.

Our efforts to capture everyday conversation behavior in speech corpora take two different directions, both of them having their own advantages and disadvantages, see [1] for a summary. The first possibility is to record the speech data directly "in the field" by equipping speakers with (head-mounted) microphones and sending them out into the world where they freely talk to other people. This approach can yield really interesting in-depth insights, particularly into the social signals of the speech code and their variation [2]. The downside of this approach is that it creates a considerable amount of heterogeneous and not consistently relevant data. In addition, unfettered long-term recordings necessarily have to focus on a few speakers, and often preclude analyzing the speech signals of the dialogue partners and their metadata.

It is certainly in part for these reasons that researchers more often try to "take the field into the laboratory" when making speech recordings. This second possibility seems easier at first sight, and maybe it actually is, but it also has a lot of pitfalls that should not be underestimated [1].

For instance, the speakers' experience with recording situations and their familiarity with dialogue partners matter a lot. Therefore, laboratory recordings should also include a practice phase or some kind of warm-up task, as well as a careful briefing before *and* after the recording. De-briefings should not be underestimated. They give useful information about how the speakers got along with the task and the language material. This can lead to excluding some speakers from further analysis, for example, because of unforeseeable interferences from non-native languages or problems with interpreting target words and cues to speech register.

Instructions in lab recordings must be standardized and checked for ambiguities. Technical equipment should be hidden away as much as possible, as it has the potential to intimidate or distract speakers. In fact, many potential pitfalls of lab recordings are not even known as yet. Who knows, for example, if and how the size of the recording booth, its surfaces, and the brightness inside affect speech production? Does eye contact with a dialogue partner affect speech production? So far, we only know that this is true for Lombard speech [3]. And who would have thought that trivial factors like daytime and the font used for read texts have significant effects on the speakers' prosodic patterns?

Kiel has a long tradition in the innovative creation and detailed analysis of speech corpora. The "Kiel Corpora of Read and Spontaneous Speech" [4] had a major influence on current models of German phonetics, phonology, and digital speech processing. Continuing this tradition, new speech corpora have been set up in Kiel - and are partly still being extended. This next generation of corpora follows the shift in research focus outlined above and constitutes the empirical foundation of the Kiel Research Center ”Speech & Emotion” (www.speechandemotion.de). Each corpus was created with a different objective. However, the corpora also supplement each other phenomenologically, and together they address four important aspects of everyday speech: 1) Emphatic and expressive speech ("KIESEL"), 2) emotional speech ("KASPAR"), 3) turn-taking behavior in dialogues ("Lindenstraße" corpus), and 4) speech in adverse conditions/noise ("SPID"). The following sections briefly outline each corpus.

The "KIESEL" Corpus

General Description

The acronym "KIESEL" stands for "KIEler Sammlung Expressiver Lesesprache" (Kiel Collection of Expressive Read Speech). The recordings stared from scripted dialogues on a wide range of everyday topics, such as holiday experiences, recent football matches, trouble at work or with the family, party planning, and annoying politicians or professors. Speakers received the straightforward instruction to produce their scripts in an informal, everyday fashion, as if the dialogues were developing spontaneously in the given setting.

In order to help the speakers manage their task, they were explicitly allowed to extemporize by inserting, omitting, or changing words or wordings (the experimenter only stepped in when this leeway affected target words, which happened in less than 1 % of all cases). Moreover, common speech reduction patterns of, for example, function words were already included in the orthography. The two speakers of a dialogue had to be good friends or in other ways very familiar with each other (e.g., relatives). They were granted as

much time as they needed prior to the recording (in practice 30-60 minutes) to get accustomed to the dialogues and the recording environment - a sound-treated room at Kiel University. In addition to being very familiar with each other, all speakers were selected to have an expressive personality and Standard Northern German as their native language.

The corpus was set up - and is still being extended - in order to study the forms and functions of emphatic accentuation in Northern Standard German. As is implied by the wide range of dialogue topics, the target categories of emphasis as well as their realization on specific target words were primarily controlled by means of the wording or, more generally, the semantic-pragmatic contexts of the dialogues. These contexts were further enriched by detailed (written) background descriptions of the situation and the acting dialogue partners. Additionally, there was a small photograph of a facial expression on top of most of the dialogues. The photograph illustrated the general mood of the dialogue situation.

The KIESEL corpus is no coherent corpus, for example, in the sense that the dialogues always had a similar length or were all read by the same set of speakers. A new set of dialogues was created and new dialogue partners were screened and recruited for each research question. Some dialogues were highly interactive and thus approximately balanced as to the amount of speech produced by each dialogue partner, whereas other research questions required eliciting almost monologues with only a few backchannels from the other speaker. Not least for this reason, the name KIESEL (pebbles) is an adequate reflection of the corpus' content. It is truly a collection of tailored quasi-spontaneous speech samples, however, held together by a joint elicitation method and aim.

Key Figures

The KIESEL corpus consists of about 4 hours of speech, produced by almost 50 speakers: 17 males and 31 females. All of them were native speakers of Northern Standard German and between 21 and 59 years old (average age was 24.9 years). Audio examples of the KIESEL corpus can be accessed by the following link: http://www.isfas.uni-kiel.de/de/linguistik/forschung/kiesel/at_download/file.

Annotation and Metadata

A part (about one quarter) of the KIESEL corpus is fully segmentally and prosodically annotated. The prosodic annotation is based on the PROLAB system, which was derived from the Kiel Intonation Model [5]. PROLAB provides an inventory of empirically grounded, phonologically distinctive pitch-accent and phrase-final intonation categories, the former of which can additionally be linked with three difference levels of perceptual prominence. The remaining three quarters of the KIESEL corpus are currently being segmentally annotated by means of the web-based Munich Automatic Segmentation System, MAUS, [6]. A corresponding manual prosodic annotation will follow.

The speech data of KIESEL are complemented by a detailed set of (anonymized) metadata of each speaker, ranging from sex and age through language background, musical experience and smoking habits to day and time of recording.

The "KASPAR" Corpus

General Description

"KASPAR" means Kiel Affective SPeech ARchive. The corpus is concerned with how different emotions surface in the speech signal, in particular with respect to dynamics and levels of acoustic energy. Emotions can cause acoustic-energy changes of up to 30 dB, thus making recordings of emotional speech a real challenge. Either loud speech passages are clipped, or soft passages are at least partially masked by the noise floor of the recording equipment. Easy and established ways to deal with this problem are either to readjust the recording gain for each speaker and emotion, or to compress the dynamic range of the speech signal. However, both approaches are no real solutions, as they make it impossible for speech researchers to directly compare acoustic-energy profiles within and across speakers and emotions.

KASPAR was created to facilitate such direct and detailed acoustic-energy analyses of emotions in speech. To that end, the recordings were conducted with two microphones whose membranes were immediately adjacent to each other, but set to very different gain factors. The sensitive microphone guaranteed the highest possible signal-to-noise ratio at soft speech passages, while the insensitive microphone recorded even the loudest shouting without any clipping artifacts. The gain difference was about 24 dB and remained unchanged during all recording sessions. Likewise, the microphone-mouth distance was kept constant at 50 cm. After the recording, complex delay estimation and amplitude processing/filtering procedures were applied to the speech material in order to arrive at analyzable one-channel sound files [7].

Five distinctive basic emotions with very different phonetic characteristics were selected: fear, anger, joy, sadness, and disgust. The emotions were elicited in combination with a set of short question and statement utterances whose semantic-pragmatic content was designed to match equally well with all emotions. The utterances were produced by trained actors/speakers, as in most other corpora on emotional speech. However, unlike in every other corpus, the emotions in KASPAR were elicited with additional support of visual and tactile stimuli, like a fireworks video or a bowl of guck. Speakers had as much time as they needed to practice the utterances and familiarize themselves with the recording environment. During the recording, the emotional utterances were not produced into nothingness, but directed towards a physically present (though only listening) interlocutor.

Key Figures

Twenty-two native speakers of Northern Standard German (all of them actors or trained speakers) produced a total of 3,580 emotional utterances. In a following step, separate perceptual and psycho-physiological experiments were conducted. While the latter experiment took indirect measurements of physical reactions, the perception experiment was based on 5AFC tasks with the emotion labels as response categories. The combined results of both experiments were to cross-check whether naive listeners were able to clearly identify the intended emotion in *all* 3,580 utterances. Utterances with ambiguous emotions were removed from the corpus. A sub-sample of 225 utterances from the KASPAR

corpus is available for download at: http://www.isfas.uni-kiel.de/de/linguistik/forschung/kaspar/at_download/file

The "Lindenstraße" Corpus

General Description

The data collection was based on the Video Task, which was specifically developed for dialogue recordings [8]. Two subjects were seated in separate rooms and watched a videotape. The two created videotapes were about 15 minutes long and consisted of scenes spliced together from episodes of a popular German TV series: The "Lindenstraße". Crucially, the two tapes were similar, but non-identical. They differed in the selection, sequential order, and completeness of scenes.

After having watched their respective video, the subjects were recorded while they tried to spot the differences between what they had seen and heard. The speakers were told that the recording was part of a psychological experiment on problem solving strategies. In this way, they were not focused as much on their own verbal behavior as they would have been, if they were told that speech data collection itself was the actual aim of the recording.

As the subjects selected for the Video Task were very familiar with each other and, moreover, emotionally attached to the presented video material, finding differences between the videos was always great fun for the dialogue partners. In addition, as the dialogue partners had an in-depth understanding of the characters and the plot of the TV series, they chatted about the video material for a long time and in a very intimate way, almost, as if they talked about friends and family. The result are dialogues consisting of vivid and humorous spontaneous speech utterances, embedded in a highly interactive and phonetically rich turn-taking structure.

Key Figures

Six dialogues were recorded from four female and two male pairs. All of them were native speakers of Northern Standard German and between 20 and 35 years old at the time of the recording. The dialogues are between 9 and 15 minutes long. Thus, the entire corpus includes about 80 minutes of speech.

The Lindenstrasse corpus is presented and distributed as Volume IV of "The Kiel Corpus of Spontaneous Speech" [4]. Audio examples of the Lindenstraße corpus can be accessed by the following link: http://www.ipds.uni-kiel.de/pub_exx/bp2001_1/Linda21.html.

Annotation and Metadata

The annotations of the Lindenstraße corpus were conducted and double-checked over almost 10 years by trained research assistants of the former Kiel Institute of Phonetics and Digital Speech Processing. As a result, the Lindenstraße annotation provides a remarkable amount of detail at multiple interlinked levels: An orthographic transliteration that additionally includes special characters for breathing, pauses, etc.; a phonetic sound segmentation in SAMPA with reference to a canonical/phonemic representation; a phonological annotation of prominence and intonation based on PROLAB; a commentary level; and an annotation of the dialogue structure with separate symbols for turn-internal and turn-final boundaries, and overlapping and non-overlapping turn transitions [9]. A separate file provides a detailed set of metadata for all 12 speakers.

The "SPID" Corpus

General Description

The "SPID" (SPontaneous In-car Dialogues) corpus allows investigating - for the first time ever - the communication between speakers inside a driving car and the Lombard effect that emerges at different driving speeds. The corpus is still being at extended. Currently, speech recordings are made in order to analyze the effects of an in-car-communication (ICC) system on speech production at different driving noise levels. ICC systems [10] are meant to improve the communication of passengers inside a car and thus help increase driving safety.

At the heart of the SPID corpus is an acoustic ambiance simulation [11]. That is, the speakers sit inside a stationary car and hear realistic driving noises of exactly this car. The acoustic simulation is further complemented by a visual (screen-based) projection of real driving situations that match with the driving noises. On this basis, the Lombard effect can be investigated under highly sophisticated and at the same time highly controlled laboratory conditions; and in addition, the noise can be entirely removed again from the signals after the recordings by means of adaptive cancellation and suppression approaches [12]. Thereby, Lombard-affected speech features like intonation, stress, and formants can be analyzed in full detail, without the corresponding measurements being distorted by background noise.

The recordings themselves were based on the Map-Task paradigm [13], which elicits spontaneous speech with a number of selected target words included (names of streets, places, persons etc.). One speaker sat in the front passenger seat and the other one behind him/her on the backseat. The two speakers had their own microphone. Recordings were made in driving simulations at 50 km/h (city) and 130 km/h (highway), as well as in a silent reference condition.

Key Figures

The corpus consists of approximately 13 hours of spontaneous dialogues whose individual durations vary depending on how long it took the dialogue partners to solve their Map Task. The speaker sample included 8 male and 8 female native speakers of Standard German. They lived for a long time in Northern Germany and were between 22 and 31 years old (average age: 26.5 years) at the time of the recording. Pairs of speakers were always of the same gender. An AV example of SPID can be accessed here: http://www.isfas.uni-kiel.de/de/linguistik/forschung/projekte/resolveuid/0887b253-41d0-4576-b3f6-03c15a92ec15

Annotation and Metadata

The corpus will soon be segmentally and prosodically annotated using the web-based Munich Automatic Segmentation System [6]. Metadata were collected and can be provided.

Summary and Conclusions

Speech communication in everyday life is rich in interactional and expressive elements. However, these elements are

anything but easy to elicit inside the laboratory, which is the only place where we can gain a reasonable degree of control over target words and phonetic context factors. Researchers have invented many different role-play, quiz, and instruction-giving scenarios whose interactive and easily diverting communication tasks are able to strike a balance between spontaneously developing everyday dialogues on the one hand, and controlled, high-quality laboratory recordings on the other. The Kiel Corpora of "Speech & Emotion" have adopted these types of tasks in the Lindenstraße corpus and in the SPID corpus. The Video Task proved to be superior to regular Map Tasks insofar as it led to a more balanced interaction between the dialogue partners, for example, with respect to speaking time and social hierarchy. The Map Task, in turn, performed better than the Video Task with respect to the frequency of occurrence of target words (names of streets, places, persons), the length of dialogues, and the coherence of their content.

The Kiel Corpora of "Speech & Emotion" take special measures to increase ecological validity. Five points are common to all corpora: (1) The use of good friends as dialogue partners; (2) the suitability screening of speakers (e.g., with respect to an expressive character); (3) the long familiarization phase prior to speech recordings; (4) the particularly rich and detailed semantic-pragmatic and situational contextualization of speech recordings; and (5) the separate line of methodology oriented research, aimed at understanding the speech differences inside and outside the laboratory, and the factors that affect speech production in the laboratory.

Research on point (5) also strengthened our corpora. It was, for instance, shown that the in-car communication simulated under laboratory conditions in SPID is qualitatively the same as in an actually driving car [11]. A large-scale perception experiment was conducted for the KASPAR corpus in order to filter out utterances whose target emotion could not be clearly identified [7]. The Video Task of the Lindenstraße corpus was found to create the same kind of phonetic accommodation phenomena between the dialogue partners as in real everyday conversation [14]. For the KIESEL corpus, we showed that the script-based quasi-spontaneous dialogues are prosodically closer to real everyday dialogues than to read-speech monologues [15], and that the elicited emphatic accentuations occur also in spontaneous speech [16].

So, when factors like font type, leeway for wording, and points (1)-(4) above are taken into account, spontaneous-speech phenomena and prepared texts - allowing the experimenter to keep tabs on what is said and how - are no longer mutually exclusive issues in speech recordings. Therefore, particularly the approach taken in the KIESEL corpus - in combination with the corresponding line of methodological research - could indicate a promising direction for creating corpora that take on the challenge to jam the phonetic richness and structural multifacetedness of everyday speech communication into the laboratory.

Literatur

[1] Niebuhr, O. & Michaud, A.: Speech data acquisition - The underestimated challenge. Kieler Arbeiten in Linguistik und Phonetik (KALIPHO) 3 (2015), 1-42

[2] Campbell, N. & Mokhtari, P.: Voice quality - The 4th prosodic parameter. Proc. 15th Int. Congress of Phonetic Sciences, Barcelona, Spain (2003), 2417-2420

[3] Cooke M., Mayo, C., & Villegas, J.: The contribution of durational and spectral changes to the Lombard speech intelligibility benefit. J. Acoust. Soc. Am. 135 (2014), 874-883

[4] Kohler, K.J., Pätzold, M., & Simpson, A.: From scenario to segment: the controlled elicitation, transliteration, segmentation and labelling of spontaneous speech. IPDS, Kiel, 1995

[5] Kohler, K.J.: Modelling prosody in spontaneous speech. In: Y. Sagisaka, N. Campbell, N. Higuchi (Eds), Computing Prosody: Computational Models for Processing Spontaneous Speech. Springer, New York, 1997

[6] Kisler, T., Schiel, F., & Sloetjes, H.: Signal processing via web services: the use case WebMAUS. Proc. Service-oriented Architectures (SOAs) for the Humanities, Hamburg, Germany (2012), 1-4

[7] Pfitzinger, H.R. & Kaernbach, C.: Amplitude and amplitude variation of emotional speech. Proc. 9th International Interspeech Conference, Brisbane, Australia (2008), 1036-1039

[8] Peters, B.: The database - The Kiel Corpus of Spontaneous Speech. AIPUK 35a (2005), 1-6

[9] Peters, B.: Form und Funktion prosodischer Grenzen im Gespräch. Ein phonetischer Beitrag zur Gesprächsforschung. SVH, Saarbrücken, 2012

[10] Lüke, C., Schmidt, G., Theiß, A., &, Withopf, J.: In-Car Communication. In: G. Schmidt, H. Abut, K. Takeda, J. Hansen (Eds), Smart Mobile In-Vehicle Systems. Springer, New York, 2014

[11] Landgraf, R. Simulating complex speech-production environments. In: O. Niebuhr, R. Skarnitzl (Eds), Tackling the Complexity of Speech. Epocha, Prague, 2015

[12] Lüke, C., Theiß, A., Schmidt, G., Niebuhr, O., & John, T.: Creation of a Lombard speech database using an acoustic ambiance simulation with loudspeakers. Proc. Digital Signal Processing for In-Vehicle Systems, Seoul, Korea (2013), 1-8

[13] Thompson, H.S., Anderson, A., Bard, E., Doherty-Sneddon, G., Newlands, A. , Sotillo, C.F.: The HCRC Map Task Corpus: Natural dialogue for speech recognition. Proc. Human Language Technology Workshop, Plainsboro, New Jersey, USA (1993), 25-30

[14] Mixdorff, H.: Qualitative analysis of prosody in task-oriented dialogs. Proc. 2nd International Conference on Speech Prosody, Nara, Japan (2004), 283-286

[15] Niebuhr, O., Bergherr, J., Huth, S., Lill, C., & Neuschulz, J.: Intonationsfragen hinterfragt - Die Vielschichtigkeit der prosodischen Unterschiede zwischen Aussage- und Fragesätzen mit deklarativer Syntax. Zeitschrift für Dialektologie u. Linguistik 77 (2010), 304-346