NEW RESOURCES AT BAS: ACOUSTIC, MULTIMODAL, LINGUISTIC

Florian Schiel, Christoph Draxler, Phil Hoole, Hans G. Tillmann
bas@phonetik.uni−muenchen.de
Bavarian Archive for Speech Signals, University of Munich, Germany
www.phonetik.uni−muenchen.de/Bas

ABSTRACT

This paper gives an overview of recent developments at the *Bavarian Archive for Speech Signals (BAS)* located at the Department of Phonetics, University of Munich: Four new speech corpora have been added to the catalogue and will be briefly described. The BAS pronunciation dictionary PHONOLEX was extended by two new types of entries: empirically detected pronunciation variants and empirically collected word entries. Details will be given about the new Verbmobil II speech resources soon available via BAS and the European SpeechDat Car data collection. Apart from traditional speech resources we'll also describe our WWWTranscribe tool for Web−based annotation and the first multi−modal speech resources produced at BAS.

1. INTRODUCTION

The *Bavarian Archive for Speech Signals (BAS)* was founded in 1995 by the *Institut für Phonetik und Sprachliche Kommunikation (IPSK)* located at the University of Munich, Germany. The main role of BAS is the collection, production, maintenance and dissemination of German speech resources to the speech community. In this respect BAS acts in close cooperation with other German speech labs, with the speech industry on national and international level, with other language−specific focal points for speech resources, and with international organizations such as *ELRA* in Europe and *LDC* in USA. After 4 years this initial profile ([1]) is changing now to include multi−modal speech resources and the evaluation of applied speech technology in the near future.

This report is intended to give the potential user of German speech resources − being either purely scientific or commercially based − an overview of the most recent developments and available resources at *BAS*. A complete description of all speech resources available at BAS can be found in the following URL: *www.phonetik.uni−muenchen.de/BAS*

The following section deals with traditional speech corpora, namely with four newly added corpora of different specifications. The third section gives a brief description of the developments within the *PHONOLEX* initiative that aims at the production of a very large German pronunciation dictionary. Part four is devoted to the newly started European funded *SpeechDat Car* project, part five gives a short introduction to the WWWTranscribe annotation tool, while the last part introduces first attempts to include multi−modal speech resources into *BAS*.

2. NEW SPEECH CORPORA

2.1. Regional Variants of German − RVG1

In close cooperation with *Lucent Technologies* and *AT&T Bell Labs* during the last three years the first part of a new corpus that covers all German−speaking areas of Europe was produced ([2]). The corpus consists of in−field−recordings of 84 utterances in four different technical qualities spoken by 500 speakers (RVG1). The recording setup was designed to model the typical user situation: office and home environment, standard hardware, no restrictions on background noise. The spoken items consists of application−oriented commands and digit strings, phonetically balanced read sentences, telephone numbers and 1 minute of spontaneous monologue. The data were validated and labeled with respect to noises and errors; the spontaneous part was transliterated according to Verbmobil standards. The selection of speakers was done in correlation to the overall demographic density of the German−speaking regions of Europe. All speakers were classified in a system of 36 fine grained and 9 broad dialectal classes. Speaker profiles with age, sex, origin, education, etc. are available with the corpus.

2.2 Verbmobil

The German automatic translation project Verbmobil is in its final phase now (VMII). BAS continues to maintain the freely available parts of the collected data. In contrast to Verbmobil I the scenarios of the recorded situations were extended and the format was re−defined for a better parseability as well as a better handling of the English and Japanese parts of the corpus. The Verbmobil I (VMI) corpus containing recorded dialogues in the Verbmobil scheduling task (8 German, 3 English CDROMs) is currently extended by the following data (estimates):

- 4 CDROMs of Japanese VM I dialogues
- approx. 10 CDROMs of German VM II dialogues
- approx. 4 CDROMs of English VM II dialogues

- approx. 5 CDROMs of Japanese VM II dialogues

Furthermore, the speaker and recording database was re–defined and standardized for all data. Pronunciation lexica for the three languages, phonemic segmentations of the German part ([4]) and other linguistic resources (such as dialog act labeling, prosodic labeling, tree banks, parts of speech tagging) will be included into the final corpus.

2.3 Strange Corpus 2 'Noises' – SC2

This corpus contains for the first time real life background noise from a production site (car maintenance). The corpus was produced as a reference and training corpus for an automatic car diagnosis system, where the user simply dictates the diagnosis of a broken engine via a local wireless telephone network into the system. The corpus contains read speech of 10 different speakers with screen prompted 'automobile diagnosis phrases' recorded under real conditions in two different car maintenance halls. The language is German. All speakers are male native Germans and have never participated in such a task before. They are all experts in the field of car diagnosis. Each speaker has spoken 800 3–7 word utterances derived from a corpus of 100 different sentences resulting in a total of 8000 utterances. The corpus was validated and labeled manually with regard to noises and speech errors.

2.4 Concatenative Synthesis Corpus – SI1000P

This corpus was especially designed to be used for high quality concatenative speech synthesis systems. It provides the speech of two professional radio announcers in studio quality covering a range of 1000 sentences from a German newspaper corpus. The signals contain the speech signal, the laryngographic signal and the time position of the glottal excitation analyzed from the laryngographic signal. A sub–corpus was manually segmented and labeled into phonemic segments; all recordings of one of the speakers were labeled prosodically (phrase boundaries and accents).

3. DICTIONARY: PHONOLEX

The PHONOLEX initiative aims to develop a very large pronunciation dictionary for standard German. The need for such a resource simply emerged from the fact that automatic speech recognition is not able to predict the usage of German compounds correctly. Furthermore, in the task of automatic analysis of very large speech corpora, the work load can be reduced by the lookup in a pronunciation dictionary for a first hypothesis of pronunciation. Although this approach might be a little bit of an overkill, the growing demand for PHONOLEX shows that practical solutions are needed in this field.

In 1998 the University of Leipzig joined the group with a new sort of data: empirically detected words from a exhaustive search in German day–to–day publications (mainly newspaper texts). This led to a decision to extend the PHONOLEX format to comply with the fact that entries from different sources may now in fact represent the same word entity. To separate the different sources a new key (OR) was introduced into the information line of each entry.

The extended format of PHONOLEX in its current version (2.x) contains the following items per entry:

- orthographic base form in LaTeX (required)
- word class (optional)
- genus (optional)
- text–phoneme–method (required)
- origin (required)
- citation form in SAM–PA ([3], required)
- zero or any number of empirically detected pronunciation variants in SAM–PA together with count, corpus and method of analysis

With version 2.1 PHONOLEX now comprises over 1.6 million entries. Optionally morpheme boundaries are marked in orthography and citation form.

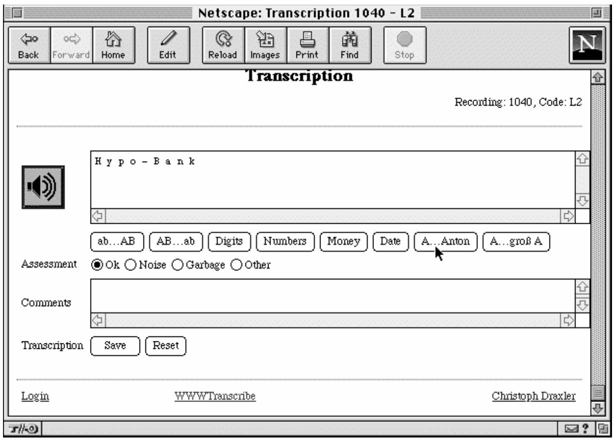
4. SPEECHDAT CAR

BAS has been involved in the SpeechDat project since its initiation: in 1994, it joined SpeechDat(M) as an associate partner to collect 1000 German speakers via the fixed network telephone; this collection was carried out in collaboration with SIEMENS AG, Germany. In SpeechDat(II), the BAS was a subcontractor to SIEMENS AG for the collection of 4000 German speakers via the fixed telephone network, and to Vocalis Ltd, UK for 1000 speakers via the mobile telephone network. Both the fixed and the mobile network database were annotated by BAS; additionally, 500 German speakers were annotated under a contract by Lernout & Hauspie, Belgium. The fixed network German database and the fixed network German databases have been validated successfully by SPEX; for the fixed network German database the error rate for speech was found to be 3.4%, the error rate for non–speech (i.e. noise) was found to be 2.0% (both values well below the limit of 5.0% and 20.0%). In the Luxemburg data the corresponding values are 4.7% 0.6%; in the mobile data 4.7% and 0.5%.

In SpeechDat–Car, BAS is responsible for the collection of the German database under a contract with Robert Bosch GmbH and BMW AG, Germany. In this project, 600 sessions will be recorded in nine languages; the recordings are carried out both in a car and synchronously via a GSM phone. In the car, four high–bandwidth channels are recorded; the vocabulary consists of application words and phrases for vehicle control, teleservices, and telecommunication commands (~ 60%), and the standard SpeechDat material (~ 40%) [11]. Finally, the BAS maintains the SpeechDat WWW server: www.speechdat.org from which all SpeechDat projects can be accessed.

4. WWWTRANSCRIBE

WWWTranscribe is a WWW–based toolbox for the orthographic annotation of speech. It was developed for SpeechDat transcriptions, but can be extended easily to other orthographic annotation systems. WWWTranscribe consists of a set of cgi–scripts written in perl that are executed by the WWW server. These scripts generate HTML formatted WWW pages containing links to speech signals and their



Netscape: Transcription 1040 – L2

Transcription

Recording: 1040, Code: L2

H y p o – B a n k

ab…AB | AB…ab | Digits | Numbers | Money | Date | A…Anton | A…groß A

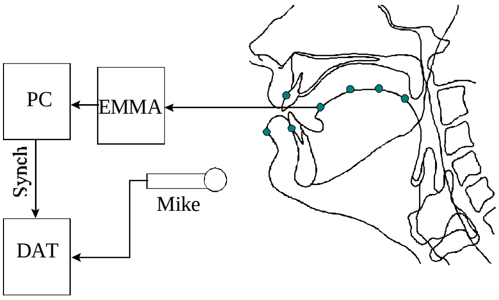
Assessment ● Ok ○ Noise ○ Garbage ○ Other

Comments

Transcription Save Reset

Login WWWTranscribe Christoph Draxler

Fig. 1 – WWWTranscribe screenshot



– Typical arrangement of sensors for EMMA experiment

corresponding transcription. The main annotation window contains a speaker button to play a speech signal, and an annotation panel with editing buttons (Fig. 1). The annotation panel contains the prompt text which is then annotated by a transcriber. Editing buttons simplify the annotation task by performing automatic conversions, e.g. from numbers to number strings, spellings to spelling alphabets, etc. Before an annotation is transmitted to the server, it is checked for formal consistency; only legal transcriptions are stored in the database.

A fully functional version of WWWTranscribe can be downloaded from the SpeechDat WWW server: www.speechdat.org/WWWTranscribe

5. MULTI–MODAL DATA

Three key categories of articulatory data have been targeted for inclusion in BAS.

5.1 Fleshpoint data

The first set of data available in this category was originally acquired for a project on vowel articulation in German. Movement data was acquired by means of electromagnetic midsagittal articulography for lower lip, jaw and four points on the tongue (see Fig. 2). High quality synchronized audio data was also acquired (16bit, 16kHz). Seven speakers spoke the following corpora:

a) a pseudo–word corpus (in a carrier phrase) of multiple repetitions of the target vowels in three consonant contexts (/p,t,k/). The corpus was recorded at normal and at fast speech rates.

b) a corpus of 105 meaningful sentences (approx. 15 syllables each) containing each target vowel in 15 different contexts.

For all corpora a manual segmentation and labeling of the target vowels is available. In addition, for the sentence material a MAUS–based segmentation ([4]) of the complete speech material has been performed.

Further details of recording techniques and articulatory pre–processing can be found in [12].

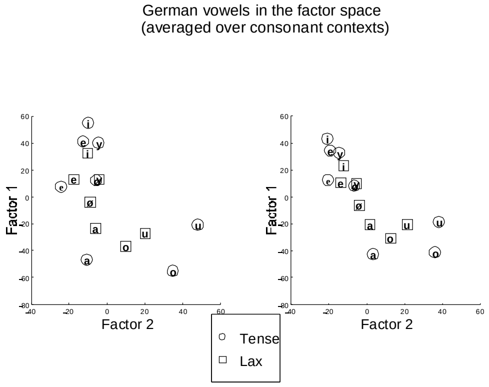
As an example of analysis results, Fig. 3 shows a speaker–independent articulatory representation of the German vowel space based on PARAFAC factor analysis of the recorded tongue configurations (further details in [14]).

Further multi–speaker corpora from more recent projects will be available in due course.

5.2. NMRI data of the vocal tract

As part of an ongoing study of alveolar consonant production, and as a supplement to the earlier vowel project just mentioned (wherever possible with the same speakers), NMRI scans have been carried out on, to date, 6 speakers producing the consonants /s/, /SH/, /l/, /n/, /t/, and the vowels /a/, /e/, /i/, /o/, /u/, /y/, /oe/. For each sound 23 slices in each of 3 planes (coronal, axial, sagittal) were collected using a T1–weighted FLASH sequence.

Basic MATLAB software for aligning the data from the different image planes and for determining vocal tract cross–sections has been developed. Coupled with EMMA data of the kind outlined above this image data should provide a substantial resource for modelling articulatory–acoustic relationships.



– Distribution of German vowels in the factor space determined from PARAFAC analysis of tongue configurations. Corpora A and B are the pseudo–word corpora (normal, fast). Corpus C is the sentence corpus.

5.3. Digitized video

We have implemented procedures for acquiring high–quality digitized video sequences in real time, together with synchronized audio (and EMMA if desired). These techniques have been used, for example, to analyze vertical larynx position in vowel production ([13]), and of course are also suitable for facial movements (e.g simultaneous frontal and profile filming using a mirror). We are currently investigating the most suitable means of preparing such very high bandwidth data (approx. 9mb/s) for further distribution.

ACKNOWLEDGMENTS

We thank all cooperating sites for their support in BAS activities. Special thanks go to the scientific staff and all the student helpers that have contributed to BAS projects, to Dr. U. Quasthoff of the University of Leipzig, to Dr. K. Netter of the DFKI Saarbrücken and to Dr. A. Wismüller of the Radiology Department, Klinikum Innenstadt, Ludwig–Maximilians–Universität, Munich, for his cooperation on the NMRI study.

REFERENCES

[1] F. Schiel (1998): Speech and Speech–Related Resources at BAS. in: Proceedings of the FIRST INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION 1998, Granada, Spain, pp. 343–349.

[2] S. Burger, F. Schiel (1998): RVG 1 – A Database for Regional Variants of Contemporary German. in: Proceedings of the FIRST INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION 1998, Granada, Spain, pp. 1083–1087.

[3] http://www.phon.ucl.ac.uk/home/sampa/home.htm

[4] A. Kipp, M.–B. Wesenick, F. Schiel (1997): Pronunciation Modeling Applied to Automatic Segmentation of Spontaneous Speech. in: Proceedings of the EUROSPEECH, Sept 1997, Rhodos, Greece, pp. 1023–1026.

[5] K. Weilhammer, S. Burger (1998): Characterizing a Database of Spoken German by Techniques of Data Mining. in: Proceedings of the FIRST INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION 1998, pp. 1253–1357, Granada, Spain.

[6] F. Schiel, S. Burger, A. Geumann, K. Weilhammer (1998): The Partitur Format at BAS. in: Proceedings of the FIRST INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION 1998, pp. 1295–1301, Granada, Spain.

[7] H. Höge, H. Tropf, R. Winski, H. van den Heuvel, R. Haeb–Umbach, K. Choukri (1997): European speech databases for telephone applications. in: Proc. of the ICASSP–97, Munich, pp. 1771–1774.

[8] C. Draxler, H. van den Heuvel, H. Tropf (1998): SpeechDat Experiences in creating large multilingual speech databases for teleservices. in: Proc. of the LREC 1998, Granada, pp 361–366.

[9] H. Höge, C. Draxler, H. van den Heuvel, F. T. Johansen, E. Sanders, H. Tropf (1999): SpeechDat Multilingual Speech Databases for Teleservices: Across the Finish Line. In: Proc. Of the Eurospeech 1999, Budapest, Hungary.

[10] C.Draxler (1997): WWWTranscribe – A Modular Transcription System Based on the World Wide Web. In: Proc. Of the Eurospeech 1997.

[11] C. Draxler, R. Grudszus, S. Euler, K. Bengler (1999): First Experiences of the German SpeechDat–Car Database Collection in Mobile Environments. in: Proc. of the Eurospeech 99.

[12] P. Hoole (1996): Issues in the acquisition, processing, reduction and parameterization of articulographic data. in: FIPKM, 34, 158–173.

[13] P. Hoole, C. Kroos (1998): ˆControl of larynx height in vowel production˜. in: Proc. 5th Int. Conf. Spoken Lang. Processing, 2, 531–534.

[14] P. Hoole (1998): Modelling tongue configuration in German vowel production. Proc. 5th Int. Conf. Spoken Lang. Processing, 5, 1863–1866.