THE PHONETIC GOALS OF THE NEW BAVARIAN ARCHIVE FOR SPEECH SIGNALS

Hans G. Tillmann, Christoph Draxler, Kurt Kotten, Florian Schiel
IPSK – Institut für Phonetik und Sprachliche Kommunikation, Munich, Germany
{tillmann,draxler,kotten,schiel}@sun1.phonetik.uni-muenchen.de
http://www.phonetik.uni-muenchen.de/

ABSTRACT

The paper describes the phonetic motivation and orientation of the new publicly funded Bavarian Archive for Speech Signals (BAS).

The BAS collects, evaluates and makes accessible very large corpora of Spoken Language (SL-) corpora. These corpora will serve to develop a (more or less) Complete Phonetic Theory (CPT) of spoken German (CPT in the sense of [8]).

INTRODUCTION

Two arguments of principal theoretical and practical interest played a major role in the decision to found a new archive for collecting spoken utterances of German, the Bavarian Archive for Speech Signals (BAS).

The first argument relates to the theoretically yet unresolved issue to what extent speakers (when producing utterances for conducting speech acts) follow the generative rules of a grammar or rather just produce copies of templates. In the latter case they have learnt to use a set of stereotypes and adopt them semantically (by selecting words from their mental lexicon) and modify them by transformation rules (such as putting in an appropiate pronominal).

How creative speakers are in making use of their language is still an open question which can only be answered by the investigation of very large corpora of empirically collected genuine speech utterances. Here, the theoretical issue of principal interest is whether the language model behind the data is best to be described by statistical or by logical (i.e. rule-based) methods.

The second argument concerns the necessity of collecting, evaluating, and making accessible very large corpora of naturally spoken utterances for the development of technical applications in the domain of Spoken Language Processing (SLP). Not only do the phonetic sciences assist and support the development of SLP technology. It is at least as important to apply SLP-methods to speech research in order to produce new phonetic knowledge.

MANAGEMENT OF VERY LARGE SL-CORPORA

SL-corpora consist of the digital speech signal and associated symbolic annotation and administration data, such as orthographic or phonemic representations of the utterances, technical specifications of the recording equipment, speaker information), or other related information.

At present, most SL-corpora are distributed on storage media, e.g. magnetic tapes or CD-ROM. Signal data is encoded in a variety of (possibly proprietary) signal file formats and symbolic data representations. All data is stored in file systems which are operating system dependent.

Clearly, this approach to SL-corpora structure and dissemination is limited:

- The size of today's SL-corpora exceeds by far the storage capacity of distribution media (TED: 7 CDs,

PhonDat: 7 CDs, Verbmobil: 6 CDs and growing).
- The lack of a standard for the representation of symbolic data leads to incompatible annotation systems, making the re-use of corpora in new contexts impossible.

SL-corpus server

A new approach to making data available consists in providing accountable network access to an *SL-corpus server.* Clients of the SL-corpus server must register to be allowed access to the data. They can either download the parts of the corpus they are interested in, or access the server on-line.

This approach has several advantages:
- Only relevant subsets of the corpus need to be accessed.
- Fine-grain control of user access to data is possible.
- Updates of a corpus become immediately available.
- Clients are shielded from storage and implementation details of the corpus data.

Network requirements

On the technical side, the SL-corpus server requires access to high-speed networks. The ISDN bandwidth of 2 x 64 Kbit/s should be considered as the lower limit and used for downloading only (uncompressed downloading a 600 MB CD will take approx. 12 hours). High-speed networks (> 100 Mbit/s) will be necessary for on-line access. Such networks are currently being deployed in Western Europe and the United States.

Data modelling requirements

On the data modelling side, the SL-corpus server requires a standard for the symbolic data representations, e.g. the computer representation of individual languages (CRIL) guidelines agreed upon at the IPA Kiel 89 convention (representation of speech data on three symbolic levels: orthography, citation form, and phonetic transcription) [2]. Furthermore, mappings of alphabets and coding systems are needed to be able to integrate data from different sources.

A first such database is the current PhonDat-Verbmobil Database, which is implemented in a persistent Prolog environment [1].

BAS technology

Since high-speed networks are not yet available everywhere, the BAS will continue the traditional dissemination of SL-corpora on CD-ROM for at least the next two years.

Currently, all SL-corpora are stored on a large central archive storage at the Leibniz-Rechenzentrum (LRZ), to which the IPSK is connected via a 100 Mbit/s fiber optic link. This archive has a capacity of approx. 2 TB (Terabyte) and a 10 GB cache.

An experimental setup to access this archive is now installed. It uses the Andrew File System (a uniform file system over multiple machines) to provide access to the data, and it supports user services such as retrieving data into the cache at a pre-specified time so that it can be accessed quickly.

A case study of using a DBMS to make available the BAS corpora is scheduled to start in summer '95.

THE WORD AS A CENTRAL PHONETIC UNIT

A narrow phonetic transcription of a human utterance is of little use if we do not know which words of which language the speaker intended to express. Even if we consider the phonetic description of single speech sounds it is important to understand that the segmental components of speech utterances could attract the interest of speech research only after alphabetic elements had become available in the form of words pronounced in phonetic citation forms [7]. Therefore the definition of the phonetic goals of the BAS was based on the decision to consider the word to be the central phonetic unit of speech research.

Single words are the first thing a new speaker of a language has to learn, and any fluent speaker of a language can easily select a word from a connected speech utterance and demonstrate it in isolation to himself and his audience. If articulated in a clear and careful pronunciation, we get citation forms which are the models for the lexical entries described in pronuciation dictionaries. Both the great stability and consistency of citation forms and the great phonetic variability of connected forms explain why the factual phonetic form of actually pronounced words remain so unconspicuous to the untrained speaker and listener. Only the clear cases represent the category.

"Les modifications phonétiques du langage" [4] had to be discoverd by the first instrumental phoneticians one hundred years ago, and they cannot be ignored by today's linguists and speech researchers because they are the true source of all the difficulties in SLP.

It is our basic theoretical assumption that the factual phonetic form of any word is a computable function of a lexically given predicate that takes a segmental structure and a prosodic shape as (a contextually independent) input. It produces an output which is context sensitive because it has to take into account the context of a prosodic phrase and a context of situation as two intervening variables.

Citation forms are computed in a specific zero context with the prosodic shape of a one-word phrase (with a terminal or an enumerating F0-contour). At the same time they also specify the functional input for computing the phonetic word form of a given connected speech utterance. Thus it makes sense to take an abstractly defined canonic citation form and relate this to the actually given pronunciations in the data base. The annotation of BAS data therefore strictly adheres to the CRIL conventions [8].

A first example of how citation forms are systematically varied to be matched to a given speech signal using an HMM-based speech verification system is described in [5].

To determine the proper predicates and the algorithms for computing the sound streams of word sequences in connected speech utterances it is necessary to be able to relate the acoustic speech signal to the articulatory production. Many purely prosodic reflexes of reduced segmental structures can only be understood if we look at individual sound gestures and their systematic reduction to allegroforms (cf. the examples given in [7]).

Therefore, multi-sensor data are of great interest and should thus be incorporated in SLP-corpora (see the final chapter for details). Concerning the relation between speech production and the digital speech signal it makes no theoretical difference whether articulatory or acoustic representations are used because today we are in a position to compute the acoustic output from the articulatory geometry.

Only if we take the word as the basic phonetic unit of speech research will we be able to understand the information-bearing variation that determines the actual form of the lexically given parts of real speech. It is the final aim of our approach to develop a theory that explains the dynamic nature of word identity in agreement with concepts such as the syllabically organized „Ausprägungscode“ proposed in [6] or of the „H and H theory“ proposed by Lindblom [3].

BAS OVERVIEW

The BAS is a publicly funded institution formally associated to the Institute of Phonetics and Speech Communication of the University of Munich.

Personnel

The permanent BAS members are Chr. Draxler who is responsible for network access and databases, K. Kotten for system administration, and F. Schiel for automatic evaluation and distribution of corpora.

Access

The BAS can be reached under the e-mail address: `bas@phonetik.uni-muenchen.de`

WWW access, including demonstrations of the corpora that are available, is possible under `http://www.phonetik.uni-muenchen.de/`

Services

The BAS provides the following services:

- Collection, evaluation, and dissemination of SL-corpora
- Customizing corpus subsets according to user specifications
- Development of corpora tools

Corpora

The BAS currently (April 95) offers the corpora listed in table 1:

Name	# Spk	# Utt	Charact.
SI1000	10	10.000	dictation
SI100	101	10.000	dictation
PhonDat I	201	16110	di-phone
PhonDat II	36	2007	train enquiry
VM 1.0.3	126	1840 turns	spontan. speech
VM 2.0	162	1538 turns	spontan. speech
TED 93	188		Eurospeech recordings, (including laryngogr)

Table 1: BAS corpora (April '95)

Corpora under development are

- VM 3.0, 4.0, and 5.0
- ERBA, a very large collection of train enquiries
- WD: phone-balanced read speech
- „Challenge Corpus", a collection of speech data that reflect problems in speech science and technology
- Polyphone-like telephone speech
- EMA, electro-magnetic articulography data of 3000 reproductions of the 15 German vowels in a defined CVC-context produced by 7 speakers in clear speech as well as in isolation

Information on these corpora can be obtained via e-mail or WWW.

REFERENCES

[1] Draxler, Chr. (1995): Introduction to the PhonDat-Verbmobil Database of Spoken German, PAP Conf. 95, Paris.

[2] IPA. (1989): The IPA Kiel Convention Workgroup 9 report: Computer coding of IPA symbols and computer representation of individual languages. *Journal of the International Phonetic Association* **19**, 81-82.

[3] Lindblom, B. (1990): Explaining phonetic variation: A sketch of the H and H theory. in: W. J. Hardcastle et al (eds): *Speech Production and Modelling*

[4] Rousselot, P.J. (1891): Les modifications phonétiques du langage, *Revue des patois gallo-romans* 4, 65-208.

[5] Schiel, F., Wesenick, M.-B. (1994): Applying Speech Verification to a Large Data Base of German to obtain a Statistical Survey about Rules of Pronunciation, *Proceedings of ICSLP 1994*, 279 - 282, Yokohama.

[6] Tillmann, H.-G. (1963), Das phonetische Silbenproblem. Phil. Diss, Bonn.

[7] Tillmann, H.-G. (1995), „Kleine und Große Phonetik", in press.

[8] Tillmann, H.-G. , Pompino-Marschall, B. (1993): Theoretical principles concerning segmentation, labelling strategies, and levels of categorical annotation for spoken language database systems. *EUROSPEECH 1993*, Berlin.