The CTH - Speech Database

An integrated multilevel approach

Per Hedelin, Dieter Huber, Per Lindblad *

Chalmers University of Technology, Department of Information Theory
* *University of Gothenburg, Department of Linguistics*
Göteborg, Sweden

INTRODUCTION

The scientific purpose of speech databases is essentially two-fold: (1) to provide the raw-material for investigative speech research, and (2) to provide the reference-material for simulative speech research.

Ideally, the same speech database serves both purposes simultaneously. In addition to that, it should also permit systematic comparison and exchange of speech data and analysis parameters between different databanks both nationally and internationally. To acheive these purposes, three basic requirements have to be fulfilled:
Standardization at all levels of data collection, registration, sampling and quantization, analysis, transcription, statistical and linguistic evaluation, database management, etc.
Integration of signal processing, analysis, and synthesis routines on the one hand, and between different levels of acoustical, statistical and linguistic analysis and evaluation on the other.
Adaptability, i.e. the speech material once collected and analysed to study one aspect of speech communication (e.g. the phonetic characteristics of certain speech sounds) should also be accessible for research on different aspects (e.g. long-time spectral properties), at different levels (e.g. prosodic variations), over different domains (e.g. complete texts), and from different points of view (i.e. auditory evaluation).

There is wide agreement in the speech research community on these points, as expressed lately during the databank meetings at the European Conference on Speech Technology (held in Edinburgh, September 1987), the Swedish Phonetics Conference (held in Uppsala, October 1987) and the Meeting of the Association for Computers and Humanities on Text Encoding Practices (held in New York, November 1987). Efforts in the direction of standardization, integration and adaptability have lately been reported among others in Itahashi [1], Ladefoged [2], and Pérennou [3].

APPLICATIONS

There are several applications of speech research that can share and use a large speech material. For speech coding it is important to have access to a variety of speakers and speaking conditions in order to evaluate the performance of a speech coder. In addition, modern speech coding is often based on large code-books, for instance for LPC-vectors. These code-books require very large training sets involving many speakers for adequate training. Essentially, the same code-books can be used in speech recognition for a raw classification of speech segments. In this case, the code-book represents the knowledge built into the recognizer regarding the mapping from phonetics to real signals. The purpose of the code-book is to split the signal space into subspaces that can be phonetically labelled. The validity of such a partitioning of the signal space, obviously, depends on the amount, and the quality, of the speech data available during training. Future development in both these areas is expected to strongly rely on the development of well trained and well balanced code-books. For speech recognition and synthesis it is vital to have a data-base for collecting information regarding, for instance, diphone data or intonation patterns. The speech material must, for these purposes, be carefully analysed and labelled from a linguistic, phonetic and prosodic point of view.

The CTH-SPEECH DATABASE

Our paper describes the approach taken at Chalmers University of Technology (Department of Information Theory) in building up an integrated multilevel speech database for the purpose of speech research (analysis and synthesis) and the development of speech coding techniques.

Given the research goals at our department, our material comprises today isolated speech sounds (phones and diphones) as well as short unrelated sentences and coherent texts. Data collection is, to start with, restricted to Swedish material and read speech. Registration of the speech samples was carried out under optimal conditions (sound insulated, unechoic studio) at the department of applied acoustics (CTH), using digital recording equipment (SONY PCM-1). Speakers were chosen in an attempt to minimize possible dialectal differences (Rikssvenska) and to represent a high standard of professionality in oral reading skills (radio journalists, experienced public speakers). The SAP signal analysis routines [4] are used for LPC, pitch estimation, inverse filtering, etc. Transcription of the entire material is performed interactively in 16 msec frames and comprises narrow phonetic transcription at the acoustical level (i.e. marking not only allophonic

variation, but even different phases in the production of various speech sounds) together with prosodic (accentuation, intonation units), sentence linguistic (constituent structure, parts-of-speech, frequency ratings) and textual analysis (text attribution, pause, breathing patterns, boundary phenomena).

THE PHONETIC LABELLING

All phonetic labelling is performed manually. Several automatic and semi-automatic methods have been considered and evaluated. With the aim of high reliability in the transcriptions in mind there were no doubts after the evaluation that manual classification is preferable.

Labelling is made in a ***sound segment*** domain. A sound segment is defined as an event in the acoustic signal. A continuant allophone consists of one sound segment. A glide has two sound segments. Stops have 4-5 sound segments. The sound segments have been chosen to ensure a close correspondence to the allophonic domain. The labelling distinguishes 52 allophones, silence included. The vowel labels are, following Elert's notation for Swedish: [ɑː], [a], [eː], [e], [ə], [iː], [ɪ], [ʊː], [ʊ], [ɯː], [ɵ], [yː], [ʏ], [oː], [ɔ], [ɛː], [ɛ], [æː], [æ], [øː], [ø], [œː] and [œ]. The consonants considered are [b], [d], [ɖ], [f], [g], [h], [j], [k], [kʰ], [l], [ɭ], [m], [n], [ɳ], [pʰ], [p], [r], [ʐ], [s], [ʂ], [ʃ], [tʰ], [t], [ʈ], [ç], [v]. Finally there is silence [ɾ], as a technical member of the allophone family.

Each of the 120 sound segment labels used is a part of a careful pronunciation of one of the 52 allophones. Labels are written in technical notation to faciliate computer handling of labels. The long vowel [iː] has the corresponding sound segment labels I11 and I12. I11 is continuant, whereas I12 corresponds to the final part in the [ij] pronunciation of /i:/. The stop [pʰ] has four corresponding sound segments, with labels P11, P12, P13 and P14. Here, P12 denotes oclusion, P13 the burst and P14 the aspiration.

The principle of using sound segments for the manual transcription allows considerable freedom. The method as such does not assume or imply that all realizations of a particular allophone use the same segments. In the actual labelling, the voice-bar of a [b], i.e. B12 can be omitted whenever needed to take one example.

SIGNAL PROCESSING

The standard processing done on all material is LPC-analysis and pitch extraction. Whenever needed, several other interactive analysis tools are available, such as spectral analysis of single frames or spectrograms of several seconds of speech. For close examination of the glottal excitation, there is an accurate and automatic inverse filtering routine.

LPC-analysis is performed using 48 ms Hamming windowed segments at an update rate of 16 ms. Adaptive pre-emphasis is employed. The auto-correlation approach to LPC is used. This is a fast and robust method that has almost no problems attached. The result of the LPC-analysis is a 10^{th}-order LPC-filter. The coefficients of the filter can be converted to the formant domain if formant estimates are required for the manual labelling.

Considerable effort has been spent on selecting an accurate and robust pitch extractor for the pitch analysis. Several different schemes have been evaluated including time domain methods such as the Gold-Rabiner algorithm, auto-correlation domain methods such as the SIFT-algorithm as well as cepstral methods. All methods tested have proven to generate errors, typically 95-99 % of the frames are correctly handled. Among the prominent problems are incorrect voicing decisions and pitch doubling. For female voices some methods occasionally give pitch estimates that are one octave below the true values. Some methods, such as the Gold-Rabiner algorithm, are more noisy in the sense that pitch estimates have low precision whereas auto-correlation methods can measure pitch with an accuracy much better than 0.5 Hz.

The final choice for the pitch extractor is an extended and improved version of the SIFT-algorithm. The original SIFT approach decimates the speech signal in order to decrease the computational burden. As an artefact, accuracy is lost in the pitch values. The basic SIFT procedure was augmented with a final analysis on the non-decimated speech signal in a vinicity of the preliminary pitch value. As a result, high precision pitch estimates were obtained with only a slight increase in computational complexity. The simple and straightforward voicing discrimination of the SIFT-algorithm was completely revised. A statistical hypothesis test was introduced. This algorithm uses speech level, spectral tilt, the normalized peak of the auto-correlation function, zero-crossing rate, rate of pitch change and formant information in order to form a composite voicing decision. In the evaluation, the pitch extractor thus designed outperformed all other methods included in the tests, in particular as regards voicing errors.

The final result of the pitch extraction contains voicing errors with typical error rates in the interval 1-2 %. Or, in other words, each sentence typically contains several incorrect voicing decisions. Manual corrections were made on a routine basis during the manual labelling step.

STATUS

The CTH-Speech Database comprises today about 2.5 hours of sampled speech data (ca 550.000 16-msec-frames). Signal processing and sentence linguistic/prosodic/textual analysis has been performed for nearly two thirds of this material (ca 320.000 frames = 1.4 hours). Narrow phonetic transcription has so far been completed for ca 42.000 frames (approx. 11 min). Further signal analysis and linguistic-prosodic-phonetic transcription is ongoing.

REFERENCES

[1] S. Itahashi (1986) A Japanese Language Speech Database, ICASSP 86, Tokyo, pp.321-324

[2] P. Ladefoged (1987) Revising the International Phonetic Alphabet, XIth ICPhS, Tallinn, Se.64.5.1

[3] G. Pérennou (1986) B.D.L.E.X.: A Data and Cognition Base of Spoken French, ICASSP 86, Tokyo, pp.325-328

[4] P. Hedelin (1986) Manual for SAP-tasks, CTH Technical Report No 5

[5] C-C. Elert (1966) Allmän och svensk fonetik, Gleerup. No 5