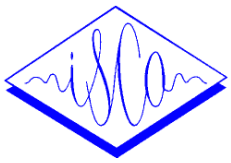

Speech Input/Output Assessment
and Speech Databases
Noordwijkerhout, the Netherlands
September 20-23, 1989

The CTH - Speech Database
An integrated multilevel approach

Per Hedelin and Dieter Huber

Department of Information Theory, Chalmers University of Technology
S-41296 Göteborg, Sweden

ABSTRACT

This paper describes the approach taken at Chalmers University of Technology in building up an integrated multilevel speech database for the purpose of speech research and the development of speech coding techniques. The material comprises today isolated speech sounds (phones and diphones) as well as short, semantically 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) using digital recording equipment (SONY PCM-F1). Segmentation, classification and labeling is performed at eight interlacing levels of linguistic (including acoustic, phonetic and prosodic) analysis.

1. INTRODUCTION

The collection and classification of large-scale speech databases constitutes one of the major activities in current speech research (e.g.[2].[4],[16],[17],[20],[23],[25],[26]) . The importance of having access to a large amount of reliably labeled speech data for both training and evaluation purposes is today not only widely accepted by most workers in the field of speech recognition, but becomes more and more generally acknowledged in the speech research community at large. The development of practical speech processing systems requires detailed knowledge as well as precise data, e.g. on the variability of phones, diphones, triphones, words and sentences with respect to different speech styles, different application domains and speaking environments, and various speaker idiosyncrasies.

Generally, the speech database provides the raw-material from which both quantitative and qualitative data on natural speech usage can be derived. However, in addition to supporting different kinds of *investigative* research, speech databases also provide the reference material for *simulative* research and for the *evaluation and assessment* of practical speech understanding systems. For instance, modern digital speech transmission systems require large speech databases for the purpose of code-book training as well as for evaluating both the objective and subjective quality of the coded speech output. In text-to-speech research the trend goes today more and more from the single-phone domain to diphone and triphone models in order to provide higher intelligibility as well as a more natural coarticulation and phonation.

Ideally, the same speech database meets the demands of both investigative and simulative research applications 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: (1) *standardization* at all levels of data collection, recording, digitization, analysis, transcription, statistical and linguistic evaluation; (2) *integration* of signal processing, analysis and synthesis routines on one hand, and between different levels of acoustical, statistical and linguistic analysis and evaluation on the other; and (3) *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 (e.g. auditory evaluation).

2. MATERIAL

Given the research goals at our department, our material comprises today isolated speech sounds (phones and diphones) as well as short, semantically unrelated sentences, and coherent texts. Systematic collection, registration and analysis is, to begin with, restricted to *Swedish* material and *read* speech. However, short samples of spontaneous, unelicited speech (both monologue and conversation) and foreign language material (American English, British English, Japanese and French) have been incorporated unsystematically into the database for domain-restricted benchmark tests. Systematic integration of spontaneous speech and non-Swedish material is under preparation for future extensions.

2.1 Phone Database

The phone database consists of (1) the 23 vowel allophones representing the nine basic vowel phonemes of Swedish [6],[8] and (2) the 100 most frequent Swedish diphones composed of CV and VC sequences [11].

2.2 Sentence Database

The sentence database contains a total of 303 comparatively short, semantically unrelated single-clause sentences that were compiled from two kinds of sources:

1 - 230 sentences were chosen from a list of phonetically balanced Swedish speech utterances [19];

2 - The remaining 73 sentences were selected from the text database described in the following paragraph, in order to study effects of list reading versus reading of coherent texts [13].

2.3 Text Database

The text database consists of five authentic Swedish newspaper articles that were selected from the *PRESS65* corpus [1] in an attempt to represent different text genres [3], different degrees of general human interest, and a statistically significant amount of the most frequent sentence structure types commonly found in Swedish [18]. Thus, one narrative (878 running words), one descriptive (730 running words),

and one argumentative text (1002 running words), together with one feuilleton (1120 running words) and one human interest story (1205 running words), comprising a total of 4935 running words, 352 graphic sentences and 102 paragraphs form the bulk of the text database.

3. SPEAKERS

The speakers employed for the recording tasks are generically divided into two groups. The four speakers in *group 1* were selected in an attempt to minimize possible dialectal differences and to represent a high standard of professionality in oral reading skills. Two of the speakers (one female, one male) are professional radio journalists working for the Swedish Broadcasting Networks *Sveriges Radio* where they are regularly engaged in newsreading sessions as well as running their own programs. The remaining two speakers (one female, one male) can be characterized as experienced public speakers, i.e. unencumbered by microphones, recording equipment and large audiences. All four were born, brought up, and are living and working in Göteborg, using the local version of Standard Swedish (Rikssvenska) both at home and at work as their normal daily vernicular.

Speakers in *group 2* come from a variety of regional, social and professional backgrounds. They were selected in an attempt to incorporate in a systematic and controlled manner the principle dialectal variations of Swedish, as well as different age groups (including children) and speech styles. No previous experience or professionality in oral reading skills was required from the subjects in this group. Up to now, recordings of twelve speakers pertaining to the second group have been incorporated in the speech database.

4. RECORDING PROCEDURES

The speech samples were recorded under optimal conditions (sound-insulated, unechoic studio) using a Brüel & Kjær 4165 microphone (complete with a Brüel & Kjær 2804 power supply) and digital recording equipment (SONY PCM-F1 audio processor combined with SONY SL-F1E video tape recorder and NAD 3020B amplifier) set to 16-bits quantization at a fixed sampling rate of 44.1 kHz. Most of the material was recorded at the department of applied acoustics at Chalmers University of Technology. Additional material, recorded under similar conditions, was received from the Swedish Telephone Company *Televerket* in Stockholm.

Only minimal instructions were given to the speakers on how to render the material. If the speaker at any time departed from his or her intended rendering of an utterance, (s)he was asked to simply utter the sentence or text passage once again. No instructions as to the placement of emphatic or contrastive stress on any word or syllable were given. During recordings, the subject and the experimenter communicated via intercom or upon the experimenter's entrance into the sound chamber between trials, but were otherwise in no visual contact with one another.

5. SIGNAL PROCESSING

For purposes of analysis and storage, the original recording on the video master tapes were down-sampled to 8 kHz (maintaining 16-bits quantization) and transfered to auxiliary disk storage. Down-sampling was acheived by digital filtering using the linear-phase transfer filter in the processing unit of the OROS.AI interface.

Standard signal processing performed on all material includes 1) filtering, 2) LPC analysis, and 3) pitch extraction.

5.1 Filtering

The Brüel & Kjær microphone (type 4165) as well as the analog pre-processing have a very low cut off-frequency. Moreover DC-drift problems were found to occur in the A/D-system. For this reason the digital processing included a high-order (512 point) FIR-filter designed to suppress frequencies below 60 Hz. As a consequence, the bandwidth of the digitized signal extends from 60 Hz to 3.95 kHz.

5.2 LPC Analysis

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 labeling.

5.3 Pitch Determination

Pitch estimates are obtained at 16 ms intervals and calculated to the first decimal. Considerable effort has been invested in designing a fast, accurate and robust pitch determination algorithm (PDA). Several approaches have been tested and systematically evaluated, including both time-domain and short-term analysis PDAs such as the Gold-Rabiner algorithm [9], the SIFT algorithm [22], and the cepstrum method [24]. The PDA finally adopted for pitch extraction constitutes an extended and improved version of the SIFT algorithm, i.e. following a basic autocorrelation approach [12].

As a result, high precision pitch estimates were obtained with only a slight increase in computational complexity. In the evaluation, the pitch extractor thus designed outperformed all other methods included in the tests, in particular with respect to incorrect voicing decisions and pitch doubling.

Remaining pitch *detection* errors were mostly of the segmentation type and were obviously caused by the incongruity between the segment boundaries and the analysis frame size. These errors were hand-edited during transcription by marking a frame as voiced if more than one third of its total duration contained a periodic speech signal. Thus an inherent segmentation error of approximately ± 10 msec has to be taken into account in all further analyses involving durational properties of the speech wave.

Pitch *estimation* errors were either of the "octave error" type, i.e. the F_0 values were wrong (or correct!) by a factor of 2 due to "jitter" in the voice source, or they were caused by irregular, aperiodic

vibrations (both in period duration and waveform) at the onset or offset of phonation. These voicing "errors" were not corrected as their detection, classification and function as potential boundary cues in speech communication is to be investigated in a series of studies conducted at our department.

5.4 Other Signal Processing

Whenever needed, several other interactive analysis tools are available from the SAP signal analysis package [10], providing for instance spectrogram displays, spectral history data both of raw input and LPC-smoothed speech, filter design and reviewing facilities, as well as a number of statistical tools for pattern classification and principle component analysis. For close examination of the glottal excitation, there is an accurate and automatic inverse filtering routine.

6. LINGUISTIC ANALYSES

The entire material is segmented, classified and labeled at eight interlacing levels of linguistic (including acoustic, phonetic and prosodic) analysis, comprising 1) acoustic event, 2) allophone, 3) phoneme, 4) word, 5) constituent, 6) clause, 7) sentence, and 8) intonation unit.

Classification and labeling at all levels of analysis is performed interactively for each 16ms-frame, with judgement based on visual evidence from oscillographic, spectrographic, and spectrum readings of the signal displayed on the graphics screen of the workstation, complemented by auditory listing.

6.1 Phonetic Labeling

All phonetic labeling (levels 1,2 and 3) is performed manually in the *acoustic event* domain (level 1). Several automatic [7],[21] and semi-automatic [5],[27] transcription methods have been considered at the outset of our database involvement. After careful evaluation, however, they have been rejected because of their potential unreliability and the prohibitive amount of manual post-editing necessary to correct faulty segmentations and/or classifications.

The ultimate objective of the *acoustic event* labeling is to delineate the signal into segments that are 1) acoustically homogeneous and 2) provide an accurate and reliable description which can help us to understand the multifarious mappings between the continuous speech signal and various sets of discrete linguistic (phonological, lexical, syntactical, etc) symbols.

A total number of 103 acoustic event labels is used to transcribe the 59 allophones representing the 27 phonemes of standard Swedish (Rikssvenska). In addition, seven non-phonological acoustic event labels are employed to describe silence (pause), breathing (speech inhalation), hesitations (vocalisations), aspiration and various kinds of phonation onset and offset phenomena (cf. [13]) commonly occurring in continuous speech utterances.

The labels are written in technical notation reflecting as far as possible Swedish orthographic spelling conventions. According to this system, the relationships between the acoustic, the allophonic and the phonemic descriptions can be directly deduced from the individual acoustic event label, where the letter (or letter combination) alone uniquely specifies the underlying phoneme, as for instance in:

AE31 - /ɛ/

the sequence consisting of the letter (or letter combination) together with the *first* digit uniquely specifies the allophone:

AE31 - [æ:]

and the second digit, finally, identifies the specific *acoustic event* within the phoneme/allophone. The number of classified acoustical events depends on the level of detail desired and may range from either "1" (for continuants such as e.g. [e] or [n]) over "1" and "2" (for glides such as e.g. [ij], where I12 corresponds to the final, semivowel-like part of the /i:/ often used in standard Swedish pronunciation) and "1" ... "3" (for instance for voiceless stops such as e.g. in [k^h], where K11 denotes the occlusion, K12 the burst and K13 the aspiration phase), to maximally "1" ... "5" (for instance in voiced stops such as [b], with B11 denoting the onset phase, B12 the voice bar, B13 a transitional phase which corresponds to the almost silent interval which sometimes occurs before the release of pressure, B14 the burst, and finally B15 the transition into the following speech sound.

The principle of using acoustic event labels for the manual transcription allows for considerable freedom and permits a description format in which information ranging from coarse to fine can be represented in an integrated and hierarchically organized fashion. The method as such, however, does not assume or imply that all acoustic realizations of a particular allophone must comprise all potentially possible acoustic events. For instance, in the actual labeling of the voiced stop [b], e.g. B12 can be omitted whenever necessary, i.e. when no occurrence of periodicity during the occlusion phase can be identified in the acoustical speech signal.

6.2 Lexical Analysis

Word identification and lexical labeling (level 4) is performed essentially by phonetic label concatenation together with top-down, expectancy driven processing guided by the analyser's knowledge of the original utterances. It must be appreciated that in view of our general objective to provide an accurate acoustic description of the actual speech signal, insertions, elisions, reductions, hesitations, false starts, repairs, and other kinds of deviation from the written texts are not counted as errors and deleted, but are fully transcribed and classified.

In addition to delineating and transcribing (in orthographic writing) each individual word as it occurs in the spoken utterance, lexical analysis also includes 1) word class or part-of-speech classification, 2) frequency rating for each lexical word based on a one-million word corpus, and 3) denomination of Swedish word accent type (acute or grave).

6.3 Syntactic Analysis

Syntactic analysis (levels 4,5,6 and 7) is performed according to the principles outlined in [13], i.e the results obtained from constituent-based syntactic analysis obtained at the department of computational

linguistics at the University of Göteborg within the SYNTAG research project are transfered to and time-aligned with the phonetically transcribed speech signal in order to establish possible interrelationships between features of text linguistic, prosodic, morpho-syntactic, lexical and phonological description on one side and acoustical events in the actual speech signal on the other. For this purpose, constituent structure is defined for up to five levels of syntactic analysis.

6.4 Prosodic Analysis

The texts in their entire lengths as well as each of the isolated sentences are segmented into prosodically defined *intonation units* reflecting the informational structure of the utterance. For this purpose, two global declination lines are computed by the *linear regression* method, which approximate the trends in time of the peaks (topline) and valleys (baseline) of F_0 across the utterance. Computation is reiterated every time the *Pearson Correlation Coefficient* drops below a preset level of acceptability. Segmentation is thus performed without prior knowledge or higher-level linguistic information, with the termination of one unit being determined by the general resetting of the intonation contour wherever in the utterance it may occur. Furthermore, once the extent of an intonation unit has been established both in the time and in the frequency domain, areas of prominence (stress) can easily be detected as overshooting or undershooting F_0 excursions.

A detailed description of the prosodic segmentation algorithm together with an evaluation of its performance on Swedish, English and Japanese speech material (both female and male) has been presented in [14]. Its application to the problem of speech parsing in continuous speech utterances is treated in [15].

7. PRESENT STATUS

The CTH speech database comprises today about 2.5 hours of sampled speech data (ca 550 000 16ms-frames). Signal processing and sentence linguistic/prosodic/textual analysis has been performed for approximately two thirds of this material (ca 320 000 frames = 1.4 hours). Phonetic labeling has so far been completed for 74 000 frames (ca 20 min). Further signal processing and linguistic-prosodic-phonetic transcription is presently advancing at an average rate of about 6000 frames (approximately 1.5 min) per week. Our medium-term objective is to complete full classification and labeling of the entire recorded material latest by the end of 1990.

Even at this intermediary stage, the CTH speech database has already been successfully employed in several research projects involving speech coding, speech input/output assessment, the development of a TTS rule synthesis system, speech recognition, neural network research, and various aspects of basic phonetic research focusing on segmental, prosodic and voice register phenomena.

REFERENCES

[1] Allén, S. (1970) Nysvensk frekvensordbok baserad på tidningstext (Frequency dictionary of present-day Swedish based on newspaper material), Almqvist & Wiksell, Stockholm

[2] Baker, J.M., D.S. Pallett & J.S. Bridle (1983) Speech recognition performance assessments and available databases, Proceedings ICASSP 83, Boston

[3] Beaugrande, R. & W. Dressler (1981) Introduction to Text Linguistics, Longman, London & New York

[4] Carré, R., R. Descout, M. Eskenazi, J. Mariani & M. Rossi (1984) The French language database: defining, planning, and recording a large database, Proceedings ICASSP 84, San Diego

[5] Collins, P. & S. Barber (1986) Fine phonetic labeling methodology for speech recognition research, Proceedings ICASSP 86, Tokyo

[6] Elert, C. (1966) Allmänn och svensk fonetik, Gleerup

[7] Elovitz, H.S., R. Johnson, A. McHugh & J.E. Shore (1976) Letter-to-sound rules for automatic translation of English text to phonetics, IEEE-ASSP, Vol.24, No.6

[8] Garlén, C. (1988) Svenskans fonologi, Studentlitteratur, Lund

[9] Gold, B. & L.R. Rabiner (1969) Parallel processing techniques for estimating pitch periods of speech in the time domain, JASA 46(2)

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

[11] Hedelin, P., D. Huber & A. Leijon (1988) Probability distribution of allophones, diphones and triphones in phonetic transcription of Swedish newspaper text, CTH Technical Report No 8

[12] Hedelin, P. & D. Huber (1990) Pitch period determination of aperiodic speech signals, submitted for presentation and publication at ICASSP 90, (forthcoming)

[13] Huber, D. (1988) Aspects of the communicative function of voice in text intonation, PhD dissertation, Göteborg/Lund

[14] Huber, D. (1989) A statistical approach to the segmentation and broad classification of continuous speech into phrase-sized information units, ICASSP 89, Glasgow

[15] Huber, D. (1989) Parsing speech for structure and prominence, Proceedings of the International Workshop on Parsing Technologies, Carnegie-Mellon University, Pittsburgh

[16] Hendriks, J.P.M. & L. Boves (1988) Definition of relations in an acoustic phonetic database, Proceedings ICASSP 88. New York

[17] Itahashi, S. (1986) A Japanese language speech database, Proceedings ICASSP 86, Tokyo

[18] Jörgensen, N. (1976) Meningsbyggnaden i talad svenska, Studentlitteratur, Lund

[19] Korsan-Bengtsen, M. (1973) Distorted speech audiometry, Acta Ota-Laryngologica, Suppl. 310

[20] Kuwabara, H., K. Takeda, Y. Sagisaka, S. Katagiri, S. Morikawa & T. Watanabe (1989) Construction of a large-scale Japanese speech database and its management system, Proceedings ICASSP 89, Glasgow

[21] Makino, S. & H. Wakita (1986) Automatic labeling system using speaker-dependent phonetic unit references, Proceedings ICASSP 86, Tokyo

[22] Markel, J.D. & A.H. Gray (1976) Linear prediction of speech, Springer Verlag, Berlin

[23] Millar, P.C., I.R. Cameron, A.J. Greaves & C.M. McPeake (1988) A very large telephone-speech database collected using a automated voice-interactive dialogue, Proceedings ICASSP 88, New York

[24] Noll, A.M. (1967) Cepstrum pitch determination, JASA 41

[25] Pérennou, G. (1986) B.D.L.E.X.: A data and cognition base of spoken French, Proceedings ICASSP 86, Tokyo

[26] Price, P., W.M. Fisher, J. Bernstein & D.S. Pallett (1988) The DARPA 1000-word resource management database for continuous speech recognition, Proceedings ICASSP 88, New York

[27] Wagner, M. (1981) Automatic labeling of continuous speech with a given phonetic transcription using dynamic programming algorithms, Proceedings ICASSP 81, Atlanta