TMH-QPSR ABSTRACTS

Speech Communication and Speech Technology

CTT-bank: A speech controlled telephone banking system - an initial evaluation

Håkan Melin, Anna Sandell and Magnus Ihse

The CTT-bank demonstration system has been created at the Centre for Speech Technology (CTT) at KTH with the purpose of displaying and evaluating speech technology in telephone banking. When calling the system, users say their name and a short digit sequence of their own choice to identify themselves. Speaker verification with a prompted four-digit password is then used to authenticate identity claims. Dialogues are finite state-based. The system asks for a single piece of information at a time, such as an account name or a money amount, and users may respond with keywords optionally embedded in carrier phrases. Three banking services have been implemented: account balance, list of recent transactions and transfer between accounts. Help is always available upon request and users may interrupt the dialogue at any time to go back to the starting point of a service loop. A small user trial with 24 subjects has been performed. Subjects enrolled in the bank and called the service up to seven times during a two-week period. A questionnaire was used to collect user opinions, and speech data was recorded and used for the evaluation of speech recognition and speaker verification components in the system. The average word error rate for speech recognition of utterances with four to seven digits was 4.6%. Concept error rates for recognition of service commands, account names, money amounts and answers to yes/no questions were 12% on average for utterances where all spoken words were represented in the lexicon. The false reject rate for the speaker verification system was 4.7% with 30 seconds of enrolment data and up to three attempts at the 2-second test utterance. Impostor attempts were not included in the trial and therefore no direct measurements of false accept rates can be made. Previous work with user trials on speaker verification in telephone banking contexts has also been reviewed. Based on additional results in those trials, it is estimated that speaker verification error rates around 1% is achievable in a telephone banking application. Combined with other means of authentication this offers a large increase in security. User opinions about the trial were positive, and the overall conclusion is that speech technology is useful in telephone banking.

TMH-QPSR, Vol. 42: 1-27

ATLAS: A generic software platform for speech technology based applications

Håkan Melin

ATLAS is a Java software library that provides a framework for building multi-lingual and multi-modal applications, especially dialogue systems, on top of speech technology components. The design is based on a layered system model, where ATLAS sits as a middleware between an application-dependent layer and the speech technology components and implements much of application-independent functionality in the system. ATLAS is itself layered with interfaces to speech technology components at the bottom and self-contained dialogue components at the top. The layered design is both efficient and flexible and is suitable for a research environment. The framework also provides support for application-dependent

layers through a structure of an application with sessions interacting with users through terminals. The terminal concept supports creating audio device-independent applications that run transparently in both telephone and desktop environments. Several speech technology components are available for use with the ATLAS framework, including text-to-speech, speech recognition and speaker verification systems. Four applications that use ATLAS have so far been developed within student and research projects at the Centre for Speech Technology (CTT), including a speech controlled telephone banking system (CTT-bank) and an automated entrance receptionist (PER).

TMH-QPSR, Vol. 42: 29-42

Swedish vowels and a new three-parameter model

Gunnar Fant

Vocal tract area functions of 13 Swedish vowels have been derived from midsagittal tracings of X-ray pictures, supported by a limited tomographic material. With a few exceptions, e.g. F2 of [u†], [o†] and [O], calculated formant frequencies show a substantial agreement with data measured during the X-ray session. The sensitivity of formant frequencies to variations in vocal tract area functions have been studied.

Observed co-variation of overall vocal tract dimensions revealed a number of dependent relationships that enable a prediction of overall length, inter-incisor distance, and asymmetries of cavity shapes from the basic specification of location and area of tongue constriction and the degree of lip-rounding. The new model thus preserves physiological constraints that make it better suited for a future adaptation to consonantal modifications than earlier three-parameter models. Nomograms of formant frequencies for systematically varied model parameters are shown. Problems related to inverse mapping, from formant frequencies to model parameters, are discussed.

TMH-QPSR, Vol. 42: 43-49

Spectral emphasis as a perceptual cue to prominence

Mattias Heldner

This paper is a first attempt at investigating whether spectral emphasis has any relevance for the perception of accented words. In particular, this study focuses on (i) whether an increase in spectral emphasis will cause accented words to be perceived as more prominent, and (ii) whether modeling of spectral emphasis in connection with accents is liable to improve the quality and naturalness of speech synthesis. A method for experimentally manipulating spectral emphasis in natural and synthesized speech without adverse effects on speech quality is proposed. This method is subsequently used to create stimuli for two perceptual experiments. The first experiment involved increasing spectral emphasis in accented words in natural speech and asking listeners to compare the prominence of the manipulated and original words. Similarly, the second experiment included comparisons of the naturalness of manipulated and original words, this time generated by an mbrola synthesis. The results of these two experiments were on the whole quite negative. Increased spectral emphasis, at least as it was implemented here, did not cause words to be perceived as more prominent. Neither did it improve the naturalness of speech synthesis. A few possible explanations of these results are discussed. However, it remains an open question whether other and perhaps more realistic implementations of increased spectral emphasis may produce more salient effects.

TMH-QPSR, Vol. 42: 51-57

T Chiba and M Kajiyama: Pioneers in speech acoustics

Gunnar Fant

It is my privilege to add some comments to the celebration of the 60-year-anniversary of the publication of the book, *The Vowel, its Nature and Structure* by T. Chiba and M- Kajiyama (1941).

This is a belated recognition of a monumental piece of work in speech acoustics which came to my attention only in the mid 1950s when I was preparing my book *Acoustic Theory of Speech Production*. It was of a great help in my processing of X-ray data of Russian speech

sounds, e.g. in deriving area functions from mid-sagital projections of the vocal tract and transverse airflow sections.

TMH-QPSR, Vol. 42: 59-60

On the speech code

Gunnar Fant

Speech technology has provided important tools for applications in man-machine communication systems and is growing rapidly. But there is a risk that expansion will be limited by insufficient attention to the potentialities of speech and language research. The symbiosis between technology and basic research that has made possible the advance, now shows a tendency to turn into polarisation. At present, speech technology is highly dependent on statistical tools and large data bases, whilst phonetic research tends to focus on narrowly defined problems or abstract issues with small or no relevance for the overall code of spoken language.

The term speech code has been coined to represent the knowledge of how linguistically defined units are realised in the speech act. In this sense it conforms with general phonetics. A more precise representation of the speech code is in programs for text-to-speech synthesis and the reverse, strategies for speech recognition from a phonetic analysis of speech wave patterns.

TMH-QPSR, Vol. 42: 61-67

Music Acoustics

Playing the turntable: An introduction to scratching

Kjetil Falkenberg Hansen

This article deals with the popular and rarely studied art form of manipulating a vinyl record by rhythmically dragging and pushing it, commonly labelled “scratching”. With sufficient practise, a Disc Jockey (DJ) can have great control over the sounds and treat the turntable as an expressive musical instrument. Even though a digital-based model of scratching might seem preferable to the vulnerable vinyl record, the acoustical behaviour of the scratch has not been formally studied until now. To gain information of this behaviour a DJ was asked to perform some typical scratching patterns. These common playing techniques and the corresponding sounds have been analysed. Since the focus of the article is on the basics of how the instrument works, an overview on standardized equipment and alternative equipment is given.

TMH-QPSR, Vol. 42: 69-79

Filename: TMHAbstracts.doc
Directory: I:\annual\Annual2001
Template: J:\tmh.kth.se\tmh\proj\common\template\tmh\Normal.dot
Title: Swedish vowels and a new three-parameter model
Subject:
Author: Personal
Keywords:
Comments:
Creation Date: 2002-06-11 16:42
Change Number: 2
Last Saved On: 2002-06-11 16:42
Last Saved By: Personal
Total Editing Time: 1 Minute
Last Printed On: 2002-06-12 10:51
As of Last Complete Printing
Number of Pages: 4
Number of Words: 1 357 (approx.)
Number of Characters: 7 737 (approx.)