DISTRIBUTED SPEECH RECOGNITION

Lukáš BURGET[1], Petr MOTLÍČEK[1],
František GRÉZL[1,2], Pratibha JAIN[2]

[1]Brno University of Technology,
Faculty of Information Technology
Božetěchova 2, 61266 Brno, Czech Republic
burget@fit.vutbr.cz ,motlicek@fit.vutbr.cz
grezl@fit.vutbr.cz

[2]OGI School of Science & Engineering, OHSU
Anthropic Signal Processing Group
20000 NW Walker Road, Bevarton, Oregon 97006, USA
pratibha@ece.ogi.edu

Abstract

This article discusses possibilities of integrating speech technology into wireless technology, allowing voice input for wireless devices. Distributed speech recognition concept and activities related to its standardization are presented. First ETSI DSR MFCC based standard is described. Work on its extension to improve robustness resulting in new standard is also presented.

Keywords

Mobile communications, Speech recognition, Robustness, Distributed speech recognition, Client-Server architecture, AURORA, ETSI

1. Introduction

In the last decades, we have seen a huge expansion of data devices communicating over wireline networks. Recently wireless communication became very popular and subsequently there is an increasing use of mobile devices, that allow to access data services even while on move. Because of limited size and portable character of these devices, we must think of alternative ways how to communicate with them. A keyboard and a mouse, used as the inputs for personal computers, are no more suitable. It is the speech that would be the ideal and natural alternative and its use as an input for controlling devices is therefore more desirable than whenever before. The speech processing and mainly speech recognition technology offer solutions for this task. We will give a brief introduction to the speech recognition and we will discus possibilities of its integration into wireless data devices. The following sections will be devoted to the distributed speech recognition and project AURORA dealing with its standardization at ETSI (the European Telecommunications Standards Institute). Two proposals of new robust DSR standards will be briefly described. The first one, which became new ETSI DSR standard, was jointly developed by Motorola Labs, France Telecom and Alcatel. The second proposal [10], which we have contributed to, was provided by Qualcomm and supported by ICSI (International Computer Science Institute, Berkeley, CA, USA) and OGI (OGI School of Science & Engineering, OHSU, OR, USA). Performance of both proposed systems is finally compared with each other and with the previous non-robust ETSI DSR standard.

2. Introduction to the Speech Recognition

Most of current systems [1], [2] for automatic speech recognition consist of three basic function blocks:

- **Feature Extraction** - In this phase, speech signal is converted into stream of feature vectors - coefficients that contain only that information about given utterance that is important for its correct recognition. Parameterization is performed to reduce the size of original speech signal data and to pre-process such signal into a form fitting requirements of following classification stage. An important property of feature extraction is the suppression of information irrelevant for correct classification, such as information about speaker (e.g. fundamental frequency) and information about transmission channel (e.g. characteristic of a microphone). Currently the most popular features are Mel frequency cepstral coefficients MFCC [3].
- **Classification** - The role of classifier is to find a mapping between sequences of speech feature vectors and recognized fundamental speech elements (words in a vocabulary, phonemes). This mapping can be done for example by simple recognizer based on Dynamic Time Warping (DTW), where the sequences of parameter vectors are stored as references. Word parameters are then compared directly with the references. More advanced classifiers are mostly based on Hidden Markov Models (HMM) [4][5], where parameters of statistical models are estimated using training utterances and their associated transcriptions. After this process, the well-trained models can be used for recognition of unknown utterances. The output of the classifier is a set of possible sequences of speech elements (hypotheses) and their probabilities.
- **Language models** - The role of language models is selection of a hypothesis sentence, which is most

likely the right sequence of speech elements of a given language. The complexity of language model depends on complexity of the problem being solved (continuous speech vs. limited number of commands). Statistical models derived from data are also often used for this purpose (N-grams). Interested reader can be referred to [12].

3. Integration of the Speech Recognition Technology in Wireless Data Devices

Recently, we could see some attempts to use speech technology for wireless devices especially for mobile phones. In most cases, all three stages of speech recognition process described above where implemented directly in device itself. The task is usually quite simple: the user can record few isolated words and connect them with some functions of the device. These functions can be then controlled by the voice. No language model is needed in this trivial case and classification is mostly based on simple DTW algorithm that induces dependency on speaker. Capabilities of state-of-art speech recognition systems are, however, much larger. Unfortunately, classification and the language modeling are very computationally expensive and memory consuming. Moreover, statistical models used in these stages are very dependent on given task and language. For these reasons, it is usually not possible to fit such a recognizer into a small portable device with limited amount of resources. In case of wireless devices, there is, however, possibility of connecting to the server providing a speech recognition for given task. Server carries out the recognition process and either returns its result in form of recognized words or directly performs appropriate action and returns requested information. Speech can be sent to the server, for example, in the form of acoustic signal over the voice channel. It has, however, several drawbacks:

- Degradation in recognition performance is observed due to coding of speech transmitted over the voice channel and due to error in channel
- There are several different systems on the market using different speech coding algorithm. Recognition system using parameters trained for one such system performs worse for another.
- Many new wireless devices other than mobile phones communicate only using data channel and they would need voice channel only for connecting to speech recognition server.

Fortunately, there is a solution overcoming these drawbacks. Feature extraction as the first stage of the recognition process can be designed very efficiently with limited resources consumption and without any dependency on the task and the language. This allows carrying out the feature extraction locally on wireless device and to transmit feature stream over the error protected data channel to the network operator, or even 3-rd party application. This application then performs remaining stages of recognition process that are already designed and optimized for a given task. This concept is called Distributed Speech Recognition (DSR).

4. Distributed Speech Recognition Standards

Standard specification is required to allow incorporation of DSR technology into wireless devices and their connection to DSR servers. First DSR standard was published by ETSI (the European Telecommunications Standards Institute) in February 2000 [8]. This standard is divided into two parts:

- **Terminal DSR Front-end** describes processing that is to be performed locally on wireless device. It includes extraction of speech features, their compression using vector quantization, formatting into data frames and their error protection by CRC. Final data frames are ready for their transition over the network.
- **Server DSR Back-end** describes processing that is to be performed on server side. It includes detection of errors in incoming data frames, mitigation of these errors and decompression of speech features that are then ready to be passed on to the recognizer. The specification of recognizer is already not part of this standard.

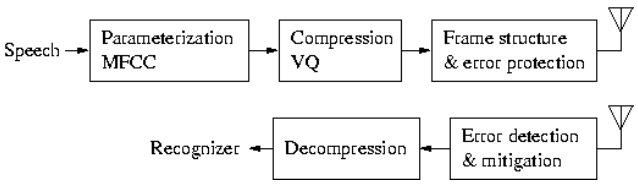

Fig. 1 Block diagram of Distributed Speech Recognition.

Feature extraction in this first standard is based on widely used Mel frequency cepstrum (MFCC) [3]. As mentioned above, the purpose of feature extraction is to reduce size of speech data and to transform these data to the form suitable for following classifier. Feature extraction based on MFCC consists of following steps:

- Speech signal is divided into segments, where the waveform can be regarded as stationary (the typical duration 25 ms). The classifiers generally assume that their input is a sequence of discrete parameter vectors, where each parameter vector represents just one such segment.
- Power Fourier spectrum is computed for every speech segment.
- Modifications inspired by physiological and psychological findings (human perception of loudness

and different sensitivity for different frequencies) are performed on spectra of each speech frame. More precisely, resolution of spectra is decreased mainly for high frequencies and log of such spectra is then taken.

- Discrete cosine transform is used to de-correlate vector for a better adaptation of features to requirements of classifier.
- Feature vectors are usually completed by first and second order derivatives of their time trajectories (delta and delta-delta coefficients). These coefficients describe changes and speed of changes of feature vectors in time.

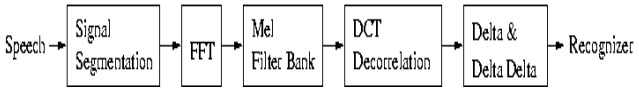

Fig. 2 Block diagram of MFCC feature extraction.

5. Robust Feature Extraction

MFCC are known to perform very well for clean input speech. However, there is a significant degradation in the performance for the real application where speech is mixed with a background noise. Mismatches in voice channel, such as differences in characteristics of microphones used by different devices, can also cause worse performance. Aurora Group at ETSI, which is responsible for DSR standardization, therefore initiated a competition that was supposed to bring new system more robust to these sources of performance degradation. To allow comparison of competing systems, new evaluation database with real-word noises was developed [7]. This database consisted of utterances with connected digits in six languages and English utterances for large vocabulary task evaluation. For the final evaluation in February 2002, there were only two competing systems proposed. The winning system that became a new DSR standard was jointly developed by Motorola Labs, France Telecom and Alcatel [11]. We have contributed to the second system proposal that was provided by Qualcomm and supported by ICSI and OGI. Both algorithms perform approximately the same, and offer about 56% reduction in error for noisy speech compared to the previous DSR standard.

6. New ETSI DSR standard

Motorola Labs, France Telecom, Alcatel system as a new ETSI DSR standard [11] use two-stages of Wiener filtering similar to conventional noise suppression technique that adaptively estimate spectrum of noise from signal parts where no speech is expected (low energy parts of signal). Speech is then processed by an adaptive filter with parameters set to filter out the portion of noise in signal. Following SNR-dependent waveform processing emphasizes portions of waveform, where signal to noise ratio (SNR) is high, and de-emphasizes low-SNR portions. MFCC features are extracted from processed speech, in the next step. Blind Equalization is finally applied to decrease dependency on channel characteristic.

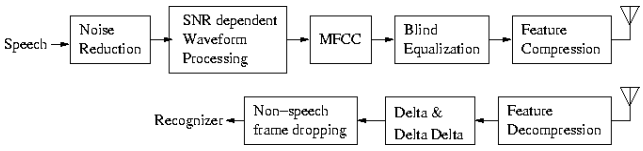

Fig. 3 Block diagram of new ETSI DSR standard.

7. Qualcomm, ICSI, OGI system

7.1 One-stream System

In the original version of Qualcomm, ICSI, OGI system, robust cepstral coefficients are computed in the terminal side using a modified Wiener filter followed by temporal filtering. A Multi-Layer Perceptron (MLP) based Voice Activity Detector (VAD) is used to detect the non-speech frames. Features are compressed using split Vector Quantization (VQ) algorithm and transmitted at a data rate of 4800 bps.

7.2 Two-stream System

The last version (Fig. 4) uses more complicated scheme at server side. Two streams of features are generated at the server from the decompressed features. The first stream (unique stream in original version) consists of cepstral coefficients that are upsampled, mean and variance normalized and augmented with first and second delta coefficients. The second stream is TempoRAl Pattern (TRAP) [9] based features. In [9], it was shown that TRAP-based features when augmented with cepstral features significantly improve the robustness. TRAP-features are based on multi-band and multi-stream approaches. For each frequency band, a feed-forward multilayer perceptron (MLP) is trained to classify speech frames. The input to each classifier is a temporal trajectory of Mel spectral energies. Fifteen Mel spectral values are reconstructed from the fifteen decompressed cepstral coefficients using Inverse DCT (IDCT). Each temporal trajectory covers a context of 50 frames in the past and 9 frames in the future. The output units of each classifier are manner-based articulatory-acoustic categories (Vowel, Flap, Stop, Fricative, Nasal) as well as the silence. The linear outputs (6 outputs) from each band-classifier are concatenated and used as input to a "merging"' feed-forward MLP. This MLP is trained to classify the same six manners of articulation targets. The mel-band MLPs as well as the merger MLP are trained using a noisy speech database, using the phoneme labeling of the database and a canonical mapping

between phonemes classes and manner classes.

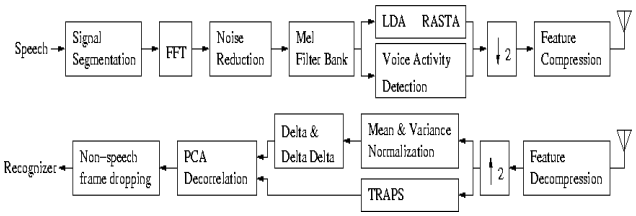

Fig. 4 Block diagram of 2-stream Qualcomm, ICSI, OGI system.

7.3 Results

Using the 2-stream system, a WER reduction of more than 58% relative to the MFCC was achieved. In Tab. 1 we compare the performance with that of the ETSI advanced front-end standard [11] using the Aurora databases. It can be seen that the latest proposed front-end is better than the ETSI advanced front-end standard for both Aurora test databases.

	QIO 1-stream	QIO 2-stream	ETSI
Aurora-2(x40%)	49.84%	55.99%	54.73%
Aurora-3(x60%)	56.62%	59.52%	56.61%
Overall	**53.91%**	**58.11%**	**55.85%**

Tab. 1 Relative improvement for QIO and ETSI front-ends for Aurora databases.

8. Summary

We have discussed possibilities of integrating speech technology into wireless communications, that allow for a new interesting voice input mainly for increasingly used small portable wireless devices. Distributed speech recognition concept as the main trend of integrating speech and wireless technology was presented together with standardization activities happening in this field. First ETSI DSR MFCC based standard was described, then systems promising new standard with improved robustness to speech in background noise were presented. Reduction in error 56% of new DSR standard compared to MFCC based DSR standard demonstrate progress happening in this area. We can expect that new portable wireless devices with sophisticated and reliable voice input will appear on the market very soon.

9. Acknowledgements

This work has been supported in part by industrial grant from Qualcomm, Inc. and in part by the Grant Agency of the Czech Republic under the project No. 102/02/0124. Pratibha Jain's stay in Brno was possible thanks to mobility grant MMI 20056 from the Czech Ministry of Education, Youth and Sports.

References

[1] Gold B., Morgan N. Speech and Audio Signal Processing. New York, 1999.

[2] L. Rabiner L., Juang B. H. Fundamentals of speech recognition. Signal Processing. Prentice Hall, Engelwood cliffs, NJ, 1993.

[3] Davis S. B., Mermelstein P. Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. on Acoustics, Speech & Signal Processing, vol. 28, no. 4, 1980, pp. 357-366.

[4] Young S. Acoustic Modeling for Large Vocabulary Continuous Speech Recognition. In Computational Models of Speech Pattern Processing, Berlin, 1999, pp. 19-39.

[5] Young S. The HTK Book. Entropics Ltd. 1999.

[6] Hermansky H., Morgan N. RASTA processing of speech. IEEE Trans. on Speech & Audio Processing, vol. 2, no. 4, 1994, pp. 578-589.

[7] Hirsch H. G., Pearce D. The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions. ISCA ITRW ASR2000, September 18-20, 2000.

[8] Pearce D. Enabling New Speech Driven Services for Mobile Devices: An overview of the ETSI standards activities for Distributed Speech Recognition Front-ends, AVIOS2000, San Jose, May 2000.

[9] Jain P., Hermansky H., Kingsbury B. Distributed Speech Recognition Using Noise-Robust MFCC and TRAPS-Estimated Manner Features. Proc. of ICSLP 2002, Denver, Colorado, September 2002.

[10] Adami A., Burget L., Dupont S., Garudadri H., Grezl F., Hermansky H., Jain P., Kajarekar S., Morgan N., Sivadas S. QUALCOMM-ICSI-OGI Features for ASR. In ICSLP, Denver, Colorado, USA, September 2002.

[11] Macho D., Mauuary L., Noe B., Cheng Y. M., Ealey D., Jouvet D., Kelleher H., Pearce D., Saadoun F. Evaluation of a Noise-robust DSR Front-end on Aurora Databases. In ICSLP, Denver, Colorado, USA, September 2002.

[12] Jelinek F. Statistical Methods for Speech Recognition. MIT Press, 1998.

About Authors...

Lukáš BURGET was born in 1975 in Vyškov, Czech Republic. He is doctoral student at the Faculty of Information Technology, University of Technology, Brno, Czech Republic. He received the Ing. degree (MSc. equivalent) in computer science from the same University in 1999. From 2000 to 2002 he was a visiting researcher at OGI Portland, USA under supervision of Prof. Hynek Hermansky. His scientific interests are in the field of speech recognition.

Petr MOTLÍČEK was born in 1976 in Brno, Czech Republic. He received the Ing. degree (MSc. equivalent) from the Department of Telecommunications, University of Technology, Brno, Czech Republic in 1999. Now he is a PhD student at the Faculty of Information Technology, at the same university. During the academic year 2000-2001 he was with ESIEE (Ecole Superieure d' Ingenieurs en Electrotechnique et Electronique) Paris. In 2001-2002 he was working in ASP group of Prof. Hynek Hermansky at OGI (Oregon Graduate Institute) in Portland, USA. His current research interests include many aspects of speech processing (coding, recognition, ...).

František GRÉZL was born in 1977 in Šternberk, Czech Republic. He is doctoral student at the Faculty of Information Technology, University of Technology, Brno, Czech Republic. He received the Ing. degree (MSc. equivalent) in electrical engineering from the same University in 2000. Since June 2001, he is visiting researcher in ASP group of Prof. Hynek Hermansky at OGI (Oregon Graduate Institute) in Portland, USA. His research interests include robust methods for speech recognition and data-driven methods for speech feature extraction.

Pratibha JAIN was born in 1973 in Jabalpur, India and she is doctoral student at OGI School of Science and Engineering, OHSU, Portland, OR, USA. She received the MSc. degree from IIT Kanpur in 1997. Since 1999, she is with the ASP group of Prof. Hynek Hermansky at OGI. In 2002, she was a summer intern at IBM T. J. Watson Center in New York. Her research interests include robustness issues in speech recognition.