

VOCALS – VOICE CENTRIC USER INTERFACES FOR LOCATION BASED SERVICES

¹Torbjørn Svendsen, ¹Andreas Egeberg, ²Trym Holter, ¹Trond Skogstad

¹Department of Electronics and Telecommunications
NTNU, 7034 Trondheim, Norway
{torbjorn, egeberg, tronds}@iet.ntnu.no

²SINTEF Information and Communication Technology
7465 Trondheim, Norway
trym.holter@sintef.no

ABSTRACT

VOCALS is a research project within the IKT2010 framework, funded by the Research Council of Norway. The project is a cooperation between the Department of Electronics and Telecommunications at the Norwegian University of Science and Technology (NTNU) and SINTEF Informatics. The project period is from 2003 to 2007.

This paper gives an overview of the VOCALS project and summarizes the most important results achieved so far.

1. Introduction

In modern society people become more and more dependent on computers and electronic communication to carry out their everyday tasks. A significant step in this development is the introduction of networked palmtop computers. This is due to a continued miniaturization of electronic circuits and deployment of a ubiquitously available broadband wireless communication infrastructure. This makes it realistic to envisage that in the near future it will be natural for people to carry with them small personal computing devices connected to the internet or a remote server via wireless communication, in the same way as mobile phones now are carried. These systems will provide access to a large set of information and communication services assisting people in whatever they have to do, also when in motion.

The handheld systems will demand a corresponding miniaturized multimodal user interface that should be able to function in an interactive manner. The most natural user interface for a small, mobile device is speech. Because of the limited size of the device it will not be feasible to use a standard type of keyboard. In addition, speech as an I/O device requires very little physical space, and is the most natural and efficient way for humans to communicate. Also, it might be operated in a

hands-/eyes-busy situation, and can even function in a multilingual mode. Thus a natural research topic is a user interface based on speech and complementary I/O-options like pen/gestures. These basic I/O-modes should be integrated with the underlying process of dialogue design.

Due to the mobility of users and the characteristics of wireless communication, the operating environment for these services will be more dynamic than what is typical for traditional distributed systems. To build systems which must operate satisfactorily under such conditions will pose new challenges and requires new solutions and new engineering methods. Services need to make the most of their surroundings and adapt themselves to overcome any limitations temporarily posed by the current user and network context, and likewise exploit any opportunities for improvement offered. For instance the service may have to adapt the user interface in response to a change in the conditions surrounding the user, or to the connection quality. Also, the service may have alternative modalities in order to fulfill the needs of a user irrespective of any disabilities (design-for-all).

It is believed that the sales of handheld, networked computers will grow rapidly and thus outnumber the sales of traditional PCs. At the same time there will be large investments in the wireless communication infrastructure. Consequently there will be a huge demand for services that utilize this infrastructure and are capable of generating sufficient income to recoup these investments. Norway, together with the other Nordic countries, is a leading region worldwide in the deployment and use of wireless communication and computers, and Norwegian software industry is in a good position to establish itself as a major player in this market. Therefore it is particularly important to build the competence needed to exploit this opportunity and disseminate it to the industry.

The VOCALS project focuses on the convergence between communication systems, advanced dialogue management and spoken language technology. This activity will be focused towards a target application; geographical information systems for mobile users. The system will be comprised by the following parts:

- A small mobile device (e.g. available in a car or as a handheld unit) with multimodal I/O-possibilities (i.e. speech, pen, text, graphics).
- A communication link between the mobile device and a remote server.
- A server containing the complete information database and most of the dialogue system modules. Parts relevant for a particular user may be downloaded to the mobile device for further processing.
- GPS (Global Positioning System) to determine the location of the user.

The foreseen application will provide the user with location-based services, e.g. tourist information about the area where the user is currently located. Information may be presented to the user by speech, text or graphics (maps). The user interface should be adaptable to the varying needs of the user. For instance, in very noisy environments a written response is preferable to a spoken one. Also, the system should be able to adapt to the resources available at a particular time.

The main goals of the project are to develop software and competence for mobile computing and speech based user interfaces and to develop a prototype for a mobile geographical information system.

2. Demonstrator prototype

The prototype under development is based on a distributed client-server architecture. The user terminal is an HP IPAQ 3970 running PocketPC 2003 and the server is a Linux PC. Communication between the server and the client is currently via WLAN. We have chosen a pilot application, which will provide tourist information for Trondheim. The client user interface will include a touch sensitive map, and the user will be able to select between spoken commands, indications of location by pen touch and pull-down menus. The prototype screen layout is shown in Figure 1.

A Point of Interest (POI) is widely defined as any clickable point in the touch sensitive map. Each POI belongs to a category, e.g. “restaurant”, “hotel” etc. Some categories will contain sub-categories, e.g. “restaurant” can contain sub-categories such as “Italian”; “Chinese” and so on. The user will interact with the system by a combination of stylus actions and spoken responses. The location of the POI is determined by the stylus position, whilst the action is defined by the spoken command. Navigation in the map can be performed by panning or zooming. The map may be zoomed in and out centered



Figure 1. Prototype client screen

around the stylus position by simple spoken commands (e.g., <action> place stylus, <say> “zoom in”).

A subset of the existing POIs will be defined as active, depending on the interaction history and/or on the visible section of the map. An example of how the system functionality is that the user touches a point in the map and asks “What is this?” The system will then show active POIs within the visible section of the map, highlight the POI nearest the stylus position and either read the name of the POI via TTS (text-to-speech synthesis) or show the name in a text window. If an appropriate POI cannot be identified (no active POIs or the stylus is determined to be too far away from any active POI), the user is notified by TTS.

The audio interface of the IPAQ will result in speech characteristics that differ from speech recorded via telephone or high quality microphones. In order to have a properly functioning speech recognizer, a small speech database will need to be collected will have to be made using the IPAQ as a recording device. We have developed a simple system for simultaneous recordings using different possible input devices, and will start collecting speech data this fall.

3. Distributed speech recognition

Internationally, much attention has been given to architectures for speech recognition on mobile platforms. In particular, it has been realized that there are three major differences between mobile speech recognition and speech recognition in other environments: i) The mobile client has very limited storage and computational capacity, ii) The transmission channel is noisy and prone to burst errors and iii) in order to optimize available band-

width, transmitted speech is rather aggressively compressed, resulting in coding distortion. On the basis of these observations, the AURORA consortium has proposed standards for distributed speech recognition (DSR) [1][2]. Distributed speech recognition is a system architecture where the relatively inexpensive speech analysis is performed in the client, and the properly error protected feature vector sequences are transmitted to a server for subsequent decoding (recognition).

An important part of the VOCALS project is to study methods for speech representation and transmission that makes the speech recognition robust to coding and transmission imperfections.

3.1 Error concealment

In the ETSI standards, a relatively simple error protection is employed, using a 4-bit CRC for each pair of feature vectors transmitted. The feature vectors consist of 13 cepstral coefficients plus log energy. Prior to transmission, the feature vectors are quantized using split vector quantization (SVQ), grouping pairs of features into 7 2-dimensional subvectors. In the ETSI standard, a frame pair is classified as erroneous if either the CRC is detected as incorrect, or if the frame pair fails a continuity test. Simple frame repetition is then applied.

In [3] it was proposed to apply a subvector-based error concealment method. Here, instead of repeating the entire feature vector if a transmission error was detected, it was attempted to identify which subvectors contained an error by applying a consistency test, and to replace these subvectors with their nearest codebook neighbors which satisfied the consistency check. This was reported to give quite promising results.

We have investigated a closely related error concealment method. The main difference to [3] is that instead of applying a spectral distance based continuity check, the subvector(s) most likely to be erroneous were identified by statistical N-grams. In the ETSI baseline [1], 6-bit codebooks are used for all subvectors except the subvector containing the log energy and zeroth cepstral coefficient, which uses an 8-bit codebook. This limits the practical length of the N-gram to bigrams (or possibly trigrams) for an exact representation of the statistics. Preliminary experiments using bigram statistics have however not shown any performance gains over the method proposed in [3]. A possible reason for this is that there may be a strong correlation between the most likely successor to a subvector and the nearest codebook neighbor satisfying the consistency check.

3.2 Feature vectors using speech coder data

For some applications it is desirable to be able to reproduce the original speech waveform. An example is automated banking applications, where approval of some transactions may require human inspection of the auditory signal.

The ETSI front-ends are not designed for good quality reproduction of the speech signal. The extended ETSI front-end [2] does include pitch and voicing information, making speech reconstruction a possibility, but as the front-end is not designed for this, the resulting speech quality is relatively low. Previous studies (e.g. [5]) have shown that if feature vectors are obtained from coded speech, either from the speech coder parameters or from the reconstructed speech, the speech recognition performance is inferior compared to feature vectors obtained from clean speech.

An alternative approach[4] would be to incorporate a speech coder at the hand held terminal (the transmitter) and to base the coding of the front-end feature vectors on the speech coded data. In addition to the coded speech, the difference between the speech coder information and the desired front-end feature vectors will be encoded and transmitted. In this way, the reconstructed speech will have the excellent quality attained by state-of-the-art speech coders while the ASR system suffer no loss in recognition rate. One set of trained models will suffice, even in scenarios where the ASR system is processing speech data that have undergone different kinds of speech coding, as the feature vectors can be made compliant with any front end processing.

The feature vector based on speech coder information can be obtained using a spectrum estimate based on either the spectrum information that most speech coders extract and transmit as side information, or by reconstructing the speech in the hand-held terminal and using that as input to the DSR feature extraction algorithm. The difference between this feature vector and the feature vector obtained by applying the DSR feature extraction to the uncoded speech can then be encoded and transmitted to the receiver. Experiments that we have performed using the 3GPP Adaptive Multi-Rate (AMR) speech coder[6], which is an Algebraic Code Excited Linear Prediction coder, have demonstrated that the difference can be more efficiently encoded when using the reconstructed speech as the basis.

Operating the AMR coder at a bit rate of 4.75 kbps, we observed we could obtain the same recognition rate as with the ETSI extended front-end using as little as 6 bits per frame to encode the difference vector. This gives a total transmission rate of 5.35 kbps, which is comparable to the source rate of 5.2 kbps of the ETSI extended front-end. I.e., at the cost of a bit rate increase of 150 bps, we can get improved speech quality and transparent speech recognition performance.

4. Cross-lingual training

Developing speech recognizers requires massive amounts of speech data for training. A common problem for smaller countries is that speech data do not exist in sufficient quantities. A particular problem that arises for

tourist information applications is that the users may have different language requirements. Having fully trained recognizers in a multitude of languages is expensive and infeasible. A possible solution to these problems is cross-lingual or multi-lingual speech recognition. In cross-lingual ASR, the idea is to utilize training data from a source language (where speech data is plentiful) to train a source recognizer, and then use the limited amount of speech data from the target language to either adapt the speech recognizer to the target language, or to obtain a mapping between the source and target language. In the VOCALS project, we are currently working on the latter approach.

The basic approach is as follows: Source language phone models are trained using a reasonably large speech database. In the target language all that is available is a fairly small speech database. In order to compensate for differences due to e.g. microphone or recording environment, a global adaptation of the source models is performed using the target data as adaptation material. We can now apply pronunciation modeling techniques in order to get a pronunciation lexicon that describes the target language in terms of the source language phones. Since the target database is fairly small, sub-word entries, e.g. syllables are used in the lexicon. We have now obtained an initial mapping between the source and the target language. However, this mapping will only apply to units seen in the target language training data. It is thus desirable to use this pronunciation lexicon as source data to obtain a more general mapping between the two languages. In order to achieve this, it is necessary to have a pronunciation lexicon for the target language (e.g. generated by the grapheme-to-phoneme module of a text-to-speech synthesizer) and a generalization algorithm.

We are currently doing initial experiments using Norwegian as a target language and Spanish as the source language. We expect to be able to report results on this work soon.

5. Conclusions

We have given an overview of the motivation and goals of the VOCALS project. A common goal of the project is to produce a prototype demonstrator of a speech centric, multi-modal location based system for tourist information. Supporting the demonstrator development are sub-projects focusing on speech recognition robustness and cross-lingual training. The project has now completed two of the four years of its planned existence, and some of the projects achievements have been presented.

6. Acknowledgements

The VOCALS project is financed by the Research Council of Norway under the IKT2010 programme.

7. References

- [1] ETSI(2003), "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms", European Telecommunications Standards Institute, Document ETSI ES201 208 v1.1.3 (2003-09), <http://www.etsi.org>.
- [2] ETSI(2003), "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Extended front-end feature extraction algorithm; Front-end feature extraction algorithm; Compression algorithms", European Telecommunications Standards Institute, Document ETSI ES202 211 v1.1.1 (2003-11), <http://www.etsi.org>
- [3] Z-H. Tan, P. Dalsgaard, B. Lindberg: "A Subvector-Based Error Concealment Algorithm for Speech Recognition over Mobile Networks", Proc. ICASSP 2004, pp. I-57-I-60, Montreal, 2004
- [4] T. Skogstad, T. Svendsen: "Distributed ASR Using Speech Coder Data for Efficient Feature Vector Representation", Proc. Interspeech 2005, Lisbon, Sept. 2005
- [5] B. T. Lilly, K. K. Paliwal: "Effect of Speech Coders on Speech Recognition Performance", Proc. ICSLP 1996
- [6] 3GPP(2004), "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Transcoding functions (Release 6)", Document 3GPP TS 26.090 V6.0.0 (2004-12), <http://www.3gpp.org>