Improving speech recognition by means of pointing gestures

María José Sánchez Martínez

Infineon Technologies AG. Corporate Research, Systems Technology. Otto-Hahn-Ring 6, 81739 Munich, Germany.
Maria.Sanchez@infineon.com

Abstract

This paper presents a multimodal system that uses the information provided by the user's pointing gestures to improve the performance of a speech recognition system. We use a Hidden Markov Models based continuous speech recognizer whose task domain is determined by an e-commerce scenario. While the user is speaking, the system keeps track of the coordinates on the screen she/he is pointing at. This is done by applying a DirectShow ®-Filter which uses a webcam to detect the user's hand movements based on skin colour. The information provided by the filter is used to rectify the recognition hypothesis when needed.

Introduction

Human-to-human dialogs are based not only on speech but also on additional information channels such as lips movements, gaze and gestures. The integration of these channels increases the robustness of the information transfer. When the dialog is held, for example, in a noisy environment, the recognition of spoken language is greatly enhanced by observing the speaker's lip movements simultaneously. Also when talking about a specific object in the room, for instance, it is often simultaneously referenced to by speech, by pointing (gestures) and by looking towards it (gaze).

Multimodal interfaces transfer the advantages of the human-to-human dialogs to the human-computer interfaces. The term multimodal refers to the ability of a system to process two or more combined user input modes (such as speech, pen, touch, manual gestures, gaze and head and body gestures) during user/system interactions [1].

The paper on hand describes our approach to a multimodal interface able to process speech and pointing input modes. We consider speech as the primary modality. The information from the pointing modality is used to improve the speech recognition's results.

This multimodal system is part of the input module of a man-machine interaction scenario. In this scenario, which will be described in detail in the next section, the user interacts with a human-like communication partner, the Virtual Personal Assistant (VPA). The VPA is able to react to the user's commands displaying natural communication forms such as speech, gesture, and mimics.

This scenario determines the task domain of the speech recognizer; the vocabulary to be recognized consists of typical phrases appearing in a sales conversation.

The Virtual Shop scenario

As a concrete application of a man-machine interaction scenario we have selected what we call a *virtual shop*. Herein, a user purchases items (in our case furniture) from an online store (see Figure1). The user communicates with the VPA, which has the form of a 3D animated talking head. This assistant presents the items, gives information on the ones selected by the user and reacts to the user's demands. The behaviour of the VPA is controlled by an advanced behavioural organization mechanism [2].

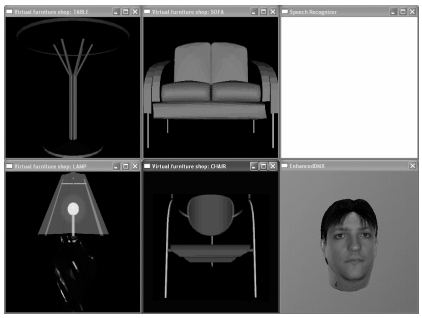

Figure 1: The Virtual-Shop scenario: a user communicates with a virtual personal assistant (VPA) on the screen (lower right) by means of natural communication channels (pointing, speech.). The VPA describes items which the user can purchase.

Multimodal Input Fusion: an Overview

The use of a multimodal system requires advanced methods to recognize each modality, and to integrate the information contained in each of them. This integration can be done at two levels: the feature level, (early integration) or the semantic level, (late fusion). Early integration applies for combinations of modalities which are closely coupled and synchronized, such as speech and lip movements. It requires, however, a large amount of training data and has high computational and training cost. Late fusion is suitable for combinations of modalities differing in the time scale characteristics of their features, like speech and pointing. In this integration strategy unimodal recognizers are used separately and are available off-the-shelf. Furthermore, modalities integration does not add any extra parameters beyond those used for each single modality recognizer. In this integration approach timing plays a crucial role. The recognition events from each modality have to be time-stamped and then integrated considering their temporal correlation.

Inputs Processing

To process speech input, we build a Hidden Markov Models based speech recognizer using ATK – the Real-Time API for HTK [3]. The recognition engine is able to generate time stamps for each word in the recognition hypotheses. It also supports a method of calculating a confidence score.

The user's pointing gestures are processed through a DirectShow ®-Filter which uses a webcam. This filter applies a simple image-processing algorithm that extracts two visual cues from the scene: colour and movement. By fusing these cues in real time the filter detects and tracks the moving hand of the user [4].

Architecture

Figure 2 shows the block diagram of our architecture.

The speech recognition module's output consists of a recognition hypothesis where each recognized word is provided along with a time stamp (time interval during which the word was spoken) and a confidence score.

While the user is speaking, the filter keeps track of the pointing cursor's x and y coordinates. In the *Virtual Shop* scenario, the screen is divided into several rectangular areas. Each of these areas is mapped to a word/group of key-words from the vocabulary. For instance, the rectangular area in the upper left corner of the screen in figure 1 is associated with the words "upper", "left" and "table" belonging to the *Virtual Shop* recognition vocabulary. By comparing the filter's output with this keywords-coordinates mapping we have a register of the words intended by the user through his/her pointing gestures.

The fusion module analyzes the output of the speech recognizer. If the confidence score of a keyword is below a certain threshold, it is replaced by the keyword supplied through the pointing module. This way, new multimodal recognition hypotheses are derived.

This architecture is implemented with a C++ based application.

Experimental Results

To test our system we use a set of ten sentences (seven words per sentence on average) from the *virtual shop* scenario's vocabulary. The transcriptions of these sentences are compared first with the speech recognition hypotheses and then with the multimodal one. Table 1 shows the results. The addition of the pointing channel allows recovering the speech recognizer's substitution errors, leading to an improvement in the recognition rate.

	Speech	Speech & Pointing
TOTAL NR. OF WORDS TO RECOGNIZE	70	70
CORRECT WORDS	59	64
DELETIONS	1	1
SUBSTITUTIONS	10	5
INSERTIONS	4	4
WORD ERROR RATE %	15,71	8,53

Table 1**:** Recognition results of the speech recognizer and the combined speech-pointing module. The test set consists of ten sentences (seven words per sentence on average) from the *virtual shop* scenario's vocabulary.

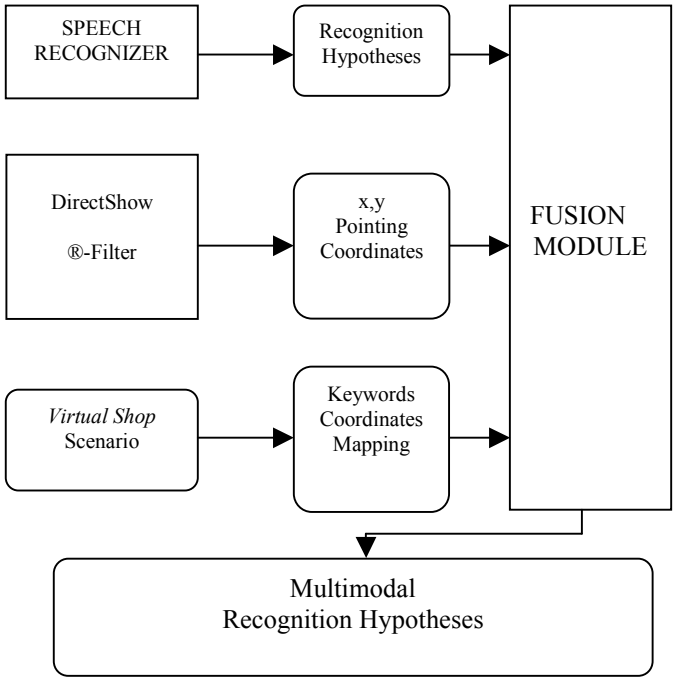

Figure 2: Architecture. The user's spoken commands are processed with a HMMs based recogniser. The recognition hypothesis includes time stamps and confidences scores of each word. The pointing gestures of the user are captured through a web cam and tracked using a DirectShow ®-Filter. The fusion module uses this information to build a new recognition hypothesis

Conclusions & Future Work

We presented an approach of a multimodal system able to integrate speech and pointing gestures. The integration of modalities is performed at semantic level. It leads to an improvement in the speech recognition results by recovering substitution errors. The system is used as the input module of a man machine interaction scenario, the *Virtual Shop.*
In further steps, the information from the user's pointing gestures will be used to solve deictic references associated to words like *this, that* or *there*

References

[1] The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications (Ed by J.Jacko & A. Sears), Lawrence Erlbaum: New Jersey, 2002

[2] D. Alonso and M.J. Sánchez, An approach to a Multimodal Man-Machine Communication System, Proceedings of the 9th International Conference "Speech and Computer" SPECOM, San Petersburg, September 2004

[3] http://htk.eng.cam.ac.uk/develop/atk.shtml

[4] Tracking Human Hand Movements by Fusing Early Visual Cues. 'Dynamic Perception', Workshop of GI section 1.0.4 "Image Understanding" and the European Networks MUHCI and ECOVISION (2002), 139-144