PROSODIC ANALYSIS OF ALERTING AND REFERENTIAL CONTEXT OF SENTINEL WORDS FOR SPONTANEOUS SPEECH INTERFACE

Veton Këpuska, Chih-Ti Shih,

Florida Institute of Technology, ECE Department
vkepuska@fit.edu

Abstract

In order to develop more natural speech-centered interfaces, it is important to consider human-to-human interaction. Humans are able to identify very quickly whether they are being addressed or not through auditory, visual, semantic and/or other contextual clues. Hence, in order to develop intelligent multi-modal user interface (MMUI) systems, it is imperative that an understanding of how humans express the need for attention be gained. Accurately predicting users' intent from speech, prosodic and visible speech features is largely an unsolved problem. In this study a model of higher level cognitive decision to discriminate between an alerting context (e.g., "Computer, Show me the chart!), and a referential context (e.g., "My computer has a quad-core Intel processor") from speech is presented. This model will serve as a basis for the proposed MMUI paradigm and it will be embedded in our present Wake-Up-Word (WUW) speech recognition (SR) system.

1. Introduction

Successful study of speech-centered MMUI will impact future computer systems, specifically by enabling development and deployment of speech-centric technologies in numerous application areas not presently possible. Those technologies will empower people with disabilities, and in general change the way humans interact with their computerized systems.

While commercial SR technologies have been around for over 25 years, they have failed to produce practical and reliable systems where humans spontaneously interact with a machine. This speech-centered user interface research goal is leading one step closer in providing the means for development of technologies that enable applications that interact with users in a natural manner. It does so by proposing a paradigm shift in the use of present SR systems. Since this paradigm is based on human-to-human interaction; it will lead to a solution that is natural and will not require special user training. Consider a scenario where a person (e.g., Jim), being part of a group of individuals, needs to draw attention of a specific person (e.g., Jill) who is out of sight. In all cases the person needed the attention (i.e., Jim) will call the other by utter the name (i.e., Jill) of that person (referred here as a sentinel word) in what we define as alerting context, for example: "Jill, I need to talk to you." In contrast, the use of the same sentinel word in other contexts is referred to as referential context: e.g., "I saw Jill today in school." In a spontaneous and natural dialog system the most important context switching occurs when a user uses alerting context. Hence, it is important to confirm that in the natural human behavior, prosodic features change in specific and deterministic ways to emphasize this semantic context. This finding will lead development of simpler algorithms to address discrimination of alerting from referential context.

In order to establish this pattern and features that capture it, we need to compare this alerting context of interest to referential context, closest yet most confusing context; acoustically/phonetically they are identical.

Recently, there has been an increasing interest in using prosody for speech technology. Unlike conventional speech processing/recognition model, prosody model provides additional information that is not directly available from transcription alone. Prosody has long been studied to provide information other than literal meaning of speech. Prosodic analysis has been applied mainly to the following four areas: (1) structural tagging such as automatic sentence segmentation of multi-party meeting 0, (2) word recognition improvement [2], (3) modeling prosodic feature sequences for speaker recognition [3] and exploiting prosodic information for speaker recognition [4] and (4) paralinguistic tagging such as emotion detection [5] and emotion-based intonation in text-to-speech system [6]. The application described in this paper uses prosodic features to determine if a sentinel word is uttered in referential or alerting context can be classified as being part of paralinguistic tagging. To the best of our knowledge the problem of discriminating between the uses of a word in an alerting context from the same word used in a referential context has never been reported.

In the section 2 importance of the alerting context in spontaneous human dialog and the architecture of general speech recognition system based on sentinel word speech recognition system (WUW-SR) is presented. Section 3 presents how prosodic structure is used by WUW-SR. In the
section 4 and 5 the results of the energy and pitch based studies are given respectively. Results combining both features are presented in section 5. Section 6 provides our plans for a additional data collection leading to a more comprehensive study and significantly more data.

2. You’re talkin’ to me?

To achieve the goal of developing a natural speech interface, it is first essential to consider human to human communication. Upon hearing an utterance of speech, a human listener must quickly make a decision whether or not the speech is directed towards him/her. Humans can make this decision quickly and robustly by utilizing auditory, visual, semantic, and/or contextual clues. Thus, understanding how humans express the need for attention is the key factor that will enable development of intelligent systems that will allow users to communicate more naturally with them. Our specific short-term goal is to enable the use of spontaneous speech as one of the primary computer interfaces - just as Captain Pickard of the Star Ship Enterprise in Star Trek series, initiates a computer interaction by referring to it as: "…Computer! <command> or <query>".

Figure 1. Software Architecture of WUW-driven general Human-Machine MMUI System.

2.1 Paradigm Shift - Sentinel Word/Phrase Recognition

To achieve the goal of spontaneous and natural speech-centric MMUI, speech recognizers must operate in continuously listening mode monitoring acoustic input all the time and must not require non-speech activation. This is in contrast to the push-to-talk model, in which speech recognition is only activated when the user pushes a button (e.g., activation of a car navigation system). Unfortunately, today’s speech recognizers are not reliable enough due to their insufficient accuracy, especially for correct rejection to operate in continuously listening mode [7]. These problems have traditionally been solved by the push-to-talk model: requesting the user to push a button immediately before talking. In the cases when this is not possible (e.g., applications over the phone) the solutions are designed with carefully crafted dialogs that limit the number of possible options at any time (e.g., mimicking touch-tone telephone keypad interface replacing it with a voice command: "Press or say I, ... ").

Hence, if the a recognition system must be listening to the user(s) all the time it needs to be capable of determining when a user verbally address the computer and not someone else. Humans use proper names to get the attention from each other explicitly. An obvious solution is to name speech recognition system with a name: a sentinel word or WUW for short that can be used to get its attention. The WUW should be recognized when spoken only in the context of requesting attention (e.g., "Computer! Dictation Mode") and should not be recognized in any other contexts including referential context (e.g., “My computer has a quad core Xeon processor.”
The WUW is analogous to the button in *push-to-talk*, but the interaction is completely based on speech allowing the user to freely interact with others when necessary.

The architecture of a speech-centric MMUI system based on the WUW-SR and WUW sentinel word recognition is presented in the Figure 1.

### 2.2 Recognition of the Alerting Context in WUW-SR

The use of a fixed sentinel word/phrase as WUW is referred to as *explicit context switching*. The second way the WUW-SR technology can be used is to have a command itself be used as WUW. This scenario is referred as *implicit context switching*.

In the voice-controlled PowerPoint presentation application the context switching example would be:

- **Explicit Context Switching**: "… as depicted from the figure in this slide, the separation of convolutional combined vocal-tract impulse response with the glottal flow in a speech signal is dependent on pitch period. In the next slide … Computer! <Next Slide> … the …"

- **Implicit Context Switching**: "… as depicted from the figure in this slide, the separation of convolutional combined vocal-tract impulse response with the glottal flow in a speech signal is dependent on pitch period. In the … <Next Slide> … the …"

The second task is more difficult for several reasons. Each command (e.g., next slide) must be recognized and not confused with other commands as well as other words, phrases, sounds, noises, etc. Also, the recognizer must discriminate each command issued in alerting context from the same WUW - sentinel word/phrase uttered in referential context. In the explicit context switching this is required only for (typically) one WUW - sentinel word/phrase (e.g., Computer). Note that in the second example above, the WUW phrase "next slide" serves dual purpose – it is referring to the content of the next slide and also it is meant to be a command that requires action from then computer. However, this dual intent of the user is very difficult to detect even for the human operator. In this work the explicit context switching will be presented.

![Figure 2. Example utterance of WUW "Wildfire" used in referential and alerting context](image-url)
3. Prosodic Structure OF Sentinel word

The WUW-SR is a highly efficient and accurate recognizer specializing in the detection of a single (sentinel) word or phrase when spoken in the alerting or WUW context of requesting attention, while rejecting all other words, phrases, sounds, noises and other acoustic events with virtually 100% accuracy when measured on all other contexts but referential. The WUW-SR is capable of discriminating limited instances of alerting contexts from referential contexts. Currently, pre-WUW and post-WUW silence is used as a prosodic feature. This feature is implemented in WUW-SR as a minimum of 50 ms leading and 75 ms trailing silence typically expected in the user’s dialog when he/she utters a sentinel word to get attention.

The waveform in the Figure 2 below represents the following sentence: “Hi. You know, I have this cool wildfire service and, you know, I'm gonna try to invoke it right now. Wildfire!”

The acoustic/prosodic clues indicated in the Figure 2 are also backed up with our intrinsic understanding of human behavior; that is, "Post-WUW" silence is expected since the person uttering this sentence is expecting a response from "Wildfire" thus silence.

The pattern depicted in the Figure 2, it is not the only way that a human would utter an attention grabbing word/phrase. For example, "Hey John, hand me that laptop!" - may have a missing "Pre-WUW" silence. Thus, the study of additional prosodic clues to expand capabilities of the current WUW-SR system as described in the next sections to achieve MMUI that enables development of context dependent speech-centric interface.

4. Prominence of Sentinel Word

To test our hypothesis that sentinel words are uttered with distinct prosodic features we investigated 12 energy based features [9] of the sentences from WUWII Corpus [7][8]. For real time speech recognition systems features that do not rely on the features after the WUW of interest are the most useful. The Figure 3 and Table 1 below depict top three results of the measurements of energy features based on five different WUWs of WUWII corpus. This result indicates that over 83% of the time WUW can be determined solely based on its prosodic structure.

![Figure 3](image)

**Figure 3.** Distribution of top three energy based features for all 5 sentinel words.

<table>
<thead>
<tr>
<th>Energy-Based Features</th>
<th>% of Samples Feature &gt; 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE1sSW_AE1SBW</td>
<td>84%</td>
</tr>
<tr>
<td>MaxEW_MaxE1SBW</td>
<td>84%</td>
</tr>
<tr>
<td>MaxE1sSW_MaxEAllIBW</td>
<td>83%</td>
</tr>
</tbody>
</table>

Table 1. Top three results based on energy feature measuring prominence of the sentinel words in WUWII corpus. AE1sSW_AE1SBW is relative change of the average energy of the first section of the WUW to the average energy of previous section just before the WUW. MaxEW_MaxE1SBW denotes relative change of the maximum energy in the WUW sections to the maximum energy in the previous section of the WUW. MaxE1sSW_MaxEAllIBW represents relative change of the maximum energy in the first section of WUW to the maximum energy in the entire speech before the WUW.

5. Melody of Sentinel word

The features based on intonation melody of an utterance containing WUW are described in this section. The melody features are computed using pitch measurements. Based on reported comparative performance studies of multiple fundamental frequency determination algorithms (FDA) [10] the Enhanced Super Resolution Fundamental Frequency Determinator (eSRFD) [11] is selected as the best algorithm of choice to perform the pitch estimation.

The pattern of the fundamental frequency contour of utterance waveforms is presented in Figure 4. We have hypothesized that alerting context of the WUW is accompanied by rise in the intonation of the WUW [9]. Based on the above hypothesis, the average and maximum pitch of segments of utterances are considered and twelve pitch-based features are derived all designed to capture the raise in the intonation of WUW. Top three pitch based features are listed in Table 2 and accompanied by the distribution of each feature plot in Figure 5.

From the Table 2, the best performance is obtained from the MaxP_MaxP1SBW - the relative change of the maximum pitch in the WUW sections to the maximum pitch in the previous section just before the WUW, is only 67%. This means that only 67% of the cases have WUW sections with higher pitch then the maximum pitch measurement in the section just before that WUW.

![Figure 4](image)

**Figure 4.** Example of the utterance of Figure 2, waveform and corresponding pitch computed using eSRFD and post-processed using median filtering.
Figure 5. Distribution of top three pitch based features for all 5 sentinel words.

Figure 6. The WUW data collection system allowing for collection of data samples as well as video for video speech analysis.

6. Conclusion

Since WUWII corpus was designed to develop acoustic models for WUWs’ over telephone channel, it is not ideal for the described study. Due to the fact that the corpus contains predominantly utterances with WUWs used in alerting context it does have data for testing the false acceptance (false negatives) obtained from utterances containing WUW’s in referential contexts. In order to be able to have explicit comparison of WUW’s in both contexts and hence develop more appropriate and discriminating prosodic features Dr. R. Wallace of UCF [12] suggested development of corpus based on extraction of audio channel from movies and TV programs from publically available DVD’s. The development of the data collection system presented in the Figure 6 below is being completed. The data collection system will enable collection of the audio data as well as video data for visual speech study.

7. References


