Cognitive State Classification in a Spoken Tutorial Dialogue System

Tong Zhang, Mark Hasegawa-Johnson and Stephen E. Levinson

Manuscript published in *Speech Communication* 48(6):616-632, 2006

Abstract

This paper addresses the manual and automatic labeling, from spontaneous speech, of a particular type of user affect that we call the *cognitive state* in a tutorial dialogue system with students of primary and early middle school ages. Our definition of the cognitive state is based on analysis of children's spontaneous speech, which is acquired during *Wizard-of-Oz* simulations of an intelligent math and physics tutor. The cognitive states of children are categorized into three classes: *confidence*, *puzzlement*, and *hesitation*. The manual labelling of cognitive states had an inter-transcriber agreement of kappa score 0.93, which was higher than strong emotion labelling in literature. For the automatic labelling, text generated by an automatic speech recognizer is searched for keyword classes and part-of-speech sequences; speech signal itself is analyzed in order to identify cepstral and prosodic correlates of cognitive states. Our study also proposes a set of cepstral features based on cognitive state-dependent speech recognition, in which the phoneme models are adapted to utterances categorized into the corresponding cognitive states. The effectiveness of the proposed method has been tested on both manually and automatically transcribed speech, and the test yielded very high correctness: 96.6% for manually transcribed speech and 95.7% for automatically recognized speech. Our study shows that the proposed cepstral features greatly outperformed the other types of features in the efficiency of cognitive state classification. Our study also shows that spectral and prosodic features derived directly from speech signals were very robust to speech recognition errors, much more than the lexical and part-of-speech based features.

Keywords: intelligent tutoring system, user affect recognition, spoken language processing.

1. Introduction[1]

1.1 A Multimodal Spoken Tutorial Dialogue System

An intelligent tutoring system (ITS) is a computer program that seeks to tutor users in an educational subject. The name *intelligent tutor* is used typically to label a system that is designed to behave as much as possible like a human tutor, rather than simply providing information and exercises like a textbook. The target application platform for the work described in this paper is an intelligent tutor in math and physics, using a Lego construction set, for children of elementary and early middle-school ages. The communication between the intelligent tutor and child users is through speech; meanwhile the intelligent tutor tracks the facial expression and activities of the users for a better estimate of the user state. The complete system has not been finished yet; the database used in this study was collected by *Wizard-of-Oz* simulations of the finished system, in which a human tutor (wizard) plays the role of the intelligent tutor. The remainder of this section describes in detail the tutorial content, the Wizard-of-Oz experiments, and the characteristics of the tutoring dialogue scenario.

1.1.1 Tutorial Content

Our intelligent tutoring system is motivated by *active learning*, also called "learning by doing." Active learning originates from *Constructionism* (Kafai and Resnick, 1996), which proposes that children can learn knowledge not solely by straightforward instruction, but also by devising some types of external artifact. Figure 1 shows a child subject playing with the Lego set.

Researchers have found that some basic mathematical concepts can be acquired through manipulating objects rather than solely handling abstract symbols (Wilensky, 1991). Therefore, we use a novel method of primary education in mathematics—playing with concrete *Lego* gears. In each experiment, the user is given gears of different sizes. We expect students to learn some basic mathematical knowledge such as *ratio* and *reciprocal* by observing the spinning activities of the Lego gears. For example, one question is about the relationship between gear size and spinning speed: *Line up a 24-tooth gear and a 40-tooth gear. If the 24-tooth gear spins 5 times, then how many times*

[1] Part of this work was presented at ICSLP 2004.

must the 40-tooth gear spin for them to line up again? Why? Children can answer this question by spinning the gears and counting the cycles. To simplify counting of cycles, the teeth on each gear are painted with a pair of different colors: red and yellow, red and blue, or yellow and blue. The traditional approach of teaching children mathematics is through memorizing formulas and rules. The proposed intelligent tutor provides children with concrete objects (Legos) to help them develop a physical understanding of some abstract mathematical concepts.

Physics is a discipline involving more direct interaction than mathematics with the physical world. The learning of some basic knowledge of physics can also be achieved by playing with Lego gears. For example, one physics question about interactive force is: *Put one hand on the 40-tooth gear axle, and put the other hand on the 8-tooth gear axle. What happens if you hold one of them steady, and try to turn the other one? Why?* Children usually think that the big gear is stronger before they do the experiment. However, it turns out that the small gear is "stronger," in the sense that it delivers greater force. Table 1 is an excerpt of the tutorial dialogue scenario.

1.1.2 Multimodal Wizard-of-Oz Experiments

System development data were acquired using a Wizard-of-Oz paradigm: children believed that they were communicating with a computer tutor instead of a human tutor, and therefore behaved as they would in a real computer-interaction environment. The system had multiple channels between student and tutor. On the one hand, the tutor used both visual objects and synthesized audio explanation to coach. On the other hand, the tutor listened to the student while tracking eye movement, body, a real-time video display of the gears, and facial expressions.

In the experiments, the user and the tutor were sitting in separate rooms. The user orally communicated with a computerized talking head shown on the computer screen ahead of him using a head-set microphone. The lip movement of the talking head was coincident with speech synthesized from text that was typed by the tutor. Identical Lego gearsets were placed in front of both the tutor and the user. Video of the location and motion of the tutor's gearset was recorded by a digital camera and displayed to the user. Video of the user's gearset and face were both simultaneously recorded by a digital camera and displayed to the tutor, and the tutor was able to listen to the user. The user's

eye movement and facial expressions were also captured by digital camera and transmitted to the tutor's computer. An instrument consisting of a receiver installed on the user's back and a transmitter carried on the head of the user was used to track the user's body position. The tutor communicated with the user by typing questions, comments, and instructions; tutor utterances were displayed on the student's computer screen and played to the user in the form of synthesized speech. Figure 2 shows the screen display of the tutor's computer and the user's computer.

1.1.3 Characteristics of the Dialogue Scenario

The tutoring dialogue scenario has three characteristics:

1. Mixed-initiative interaction—Practice provides an efficient way for children to think and motivate new ideas by themselves. Therefore, in the experiments children take the initiative to discover and learn by themselves, while the tutor serves auxiliary functions, e.g., suggesting tasks and answering student questions. The tutor usually does not over-ride the wishes of students, but the tutor initiates communication when he finds that students need guidance, assistance or encouragement. Therefore, in the dialogue system the interaction is mixed initiative. For example, the tutor can ask elicitation questions to help students better understand the knowledge involved in the phenomenon they observe, or to suggest new actions for students to perform when a task is finished. Conversely, students can ask questions if they have trouble solving a puzzle, or if they fail to understand a system utterance, or even if they dislike a puzzle that the tutor has set for them, and want to move on to something else.
2. Open-ended speech—Students are allowed to express their views unreservedly although the dialogues are in a limited domain. Educational psychologists find that it is hard for us to keep children interested in the experiments if we constrain their views and expression. So we encourage children to participate in the experiments and instigate their interests in scientific learning by asking them open-ended questions rather than single-choice questions. For example, the tutor prefers to ask "What are you noticing?" instead of "Which gear was turning faster?" Similar open-ended question examples are shown in the excerpt in Table 1, such as (1) *What are you exploring there?* (2) *What if we try just using 3 gears? What do you*

notice? (3) *What effect does the medium gear have?* (4) *Why is the smaller gear stronger?*

3. Inexperienced users—Adult users of a typical telephone-based dialogue system (e.g., for purchase of air travel, train travel, or financial instruments) are usually able to learn, over a number of repeated interactions with the system, what actions are possible at each stage of the dialogue. By contrast, children users of our intelligent tutor are inexperienced with the content of the tutoring session. Each child participates in at most three sessions, and we do not ask children to relearn knowledge that they have acquired. Therefore, the children users are always inexperienced in the content of the lessons, even though they have gained some relevant skills by using the system in the past.

1.2 User Affect in the Tutorial Dialogue System

The internal state of a talker, including such factors as strong emotion (e.g., *angry, happy, fear*), attention (e.g., *relaxed, stressed, interested, bored*), and attitude (e.g., *formal, friendly, impatient*), and information concerning the subject matter and the situation may be collectively referred to as *affect* (Gobl and Chasaide, 2003). Our preliminary analysis of the ITS corpus indicates that children do not exhibit strong emotions such as *happy* and *angry*. They also usually maintain great interest and attention during the experiments, so it is not usually possible (or necessary) to distinguish the attentional states such as *interested* vs. *bored*, and attitude such as *friendly* vs. *hostile*. Instead, their speech carries information closely related to the students' cognitive activities during the process of knowledge acquisition. Therefore, we call them *cognitive states*. For example, children's various answers to a question reflect the levels of their certainty. The questions they put forward reflect the fact that they have confusion.

We categorize the cognitive states of students into three classes: *confidence*, *puzzlement*, and *hesitation*. Generally, we classify the cognitive state to be *confidence* when the user answers questions or explains actions in relatively fluent speech (e.g., *I was looking how many times the small gear goes around the medium and then the medium goes around the large and see if they end up the same*), or issues commands (e.g., *I lost count, let's try again*), or explains the playing actions (e.g., *I'm gonna move the two 24 gears close together*), or expresses completion of a task (e.g., *Alright, I'm*

done with that). We classify the cognitive state to be *puzzlement* when the user asks questions (e.g., *How many times do both of the small tooth gears go around the medium size gear?*), or states with certainty the lack of knowledge (e.g., *I have no idea why this works*). We classify the cognitive state to be *hesitation* when the user answers questions or explains actions in heavily dysfluent and relatively slow speech (e.g., *Trains, oh, yeah, I mean on trains they have the…*), or state uncertainty (e.g., *I'm not sure*).

Our proposed cognitive states classification is not only useful for friendly user interface design (as a type of affective computing), but also useful for detecting learning activities of students and thereafter selecting appropriate tutorial tactics. In this study, the learning activities of students are summarized by being categorized using a list of about 30 application-dependent classes, where the classes are defined in order to summarize: (1) the application-dependent topic of the user's question; and (2) the implied state of the user's knowledge about the topic addressed. For example, we classify questions requesting instructions for playing with Legos into *AskForPlayInstruction*, classify those answers irrelevant to the experiment content into *IrrelevantAnswer*, and classify the utterances talking about the spinning speed into *SpinSpeed*. Many learning activities are typically associated with particular cognitive states. For example, *hesitation* is more likely to accompany *IncompleteAnswerOnSpin* (e.g., *I saw the ... yellow part ...*) or *WrongAnswerOnSpin* (e.g., *It ... I think it goes around ... one and a half times*) than *CorrectAnswerOnSpin* (e.g., *The large gear has five times as many teeth as the small ones*); *puzzlement* is a strong indicator of a question such as *AskForPlayInstruction*.

In this study, we intend to examine: (1) the efficiency of manual and automatic labeling of the cognitive states, i.e., *confidence*, *puzzlement*, and *hesitation*, that we define in a tutorial dialogue system; (2) speech characteristics such as spectrum, syllabic rate and pitch, and language characteristics such as word choice and word combination, that may reflect children students' cognitive states; and (3) the robustness of speech and language characteristics to speech recognition errors.

2. Related Work

2.1 Intelligent Tutors

Automatic intelligent tutors have been a topic of research and development for many years, and have been used for instructional purposes in many academic disciplines. For example, an intelligent tutor has been used for teaching context-free grammar through instruction and remediation (Reyes et al., 2000), an intelligent tutor has been used for teaching microscopic diagnosis by viewing virtual pathology slides (Crowley et al., 2003), and an intelligent tutor has been used for apprenticing Smalltalk programmers (Alpert et al., 1999). Some intelligent tutors have been shown to accelerate the learning efficiency of students. For example, an Air Force electronics troubleshooting tutor allowed students to gain proficiency in two weeks equivalent to the proficiency of 48-month trainees without ITS (Lesgold et al., 1990). A LISP tutor improved the speed and quality of students' programming in comparison with traditional classroom training (Corbett and Anderson, 1992). Aimed at helping college students master basic computer techniques, the AutoTutor system (Graesser, 2001) showed that students can improve their learning by 0.5 standard deviation units compared with learning by reading alone. In addition, it has even been proposed that ITS has advantages over classroom instruction, in particular, the ITS educational environment is not competitive, and thus benefits the students with special needs. For example Discover, an intelligent tutor, taught students with learning difficulties to solve mathematical word problems with less failure or frustration (Steele and Steele, 1999).

Most tutorial dialogue systems rely on text input from students, but a few tutoring systems have experimented with the use of speech in tutorial dialogues. CLT (Colorado Literacy Tutor) applies spoken dialogue system techniques in CU Communicator (Wand and Pellom, 1999; Pellom et al., 2000) to teach vocabulary to children with hearing problems and autism spectrum disorders (Cole et al., 2003). Reading Tutor is an intelligent tutor used to anticipate, detect, and remediate the difficulties of students in reading (Mostow et al., 2002; Beck et al., 2003). ITSPOKE is a spoken tutorial system that engages students in qualitative physics learning by providing feedback and correcting misconceptions, and eliciting more complete explanations (Litman and Silliman, 2004). SCoT is a spoken conversational tutor that utilizes speech recognition

and natural language understanding techniques to identify and address students' misconceptions for a shipboard damage control simulator (Clark et al., 2001; Pon-Barry et al., 2004; Schultz et al., 2003). A recent study comparing speech-based versus text-based tutorial found that speech has advantages over text in the learning gains and time required to accomplish a task when the tutor was human, but little difference between the two kinds of tutorial strategies when the tutor was a computer (Litman et al., 2004). This experimental result suggests that a perfect spoken tutorial would be more efficient than a typed tutorial, but that state-of-the-art speech techniques are not mature enough to guarantee that the spoken tutorial outperforms the typed tutorial.

2.2 Affective Computing in Spoken Dialogues

Speech conveys a variety of affective information that can be used to improve the naturalness and friendliness of user interface for human computer communication. Behavioral results show that an agent aware of user affect is significantly more effective than a neutral agent in helping relieve frustration levels. For example, in an automatic call center it is very useful to detect user affect such as *frustration*, *irritation* or *impatience*, so that a call can be redirected to a human attendant in time (Lee and Narayanan, 2005; Petrushin, 1999; Batliner et al., 2003). Many breakdowns in a telephone-based information system could be avoided if the machine was able to recognize the emotional state of the user, such as *annoyance* and *frustration*, and responded to it more sensitively (Martinovsky and Traum, 2003; Ang et al., 2003). Fernandez and Picard (2003) classify driver's speech into four levels of stress in terms of driving speed and frequency of proposing arithmetic questions. A tutorial dialogue system classifies user emotion into *positive*, *neutral*, and *negative* (Forbes-Riley and Litman, 2004) for better user modeling.

Affective activity causes physiological variations in the vocal mechanism, which is used to generate sound and causes further speech variation. The audio speech waveform carries various kinds of information that reveal user affect: long-term prosody, short-term spectrum, lexicon, syntax, and implications that carry meaning because of their relationship to dialogue context. (1) Prosody revealing phonetic variations is most commonly used for user affect recognition. Pitch is the most relevant acoustic parameter for the detection of emotion (Mozziconacci and Hermes, 1998; Juang and Furui, 2000;

Petrushin, 2000; Kang et al., 2000). For example, aroused emotions (such as *fright* and *elation*) are correlated with relatively high pitch, while relaxed emotions (such as *boredom* and *sadness*) are correlated with relatively low pitch. The other prosodic features are energy, duration, and speaking rate. (2) Certain word choices and word combinations are more likely to express certain affect. Lexical information has been encoded, for the purpose of affect recognition, in the form of affect-dependent unigram and bigram language model statistics (Batliner, et al, 2000; Polzin and Waibel, 2000). Lee and Narayanan (2005) used word-emotion mutual information to calculate emotional salience to automatically detect a set of emotionally salient words. (3) Discourse provides a knowledge source to help reveal emotion (Batliner et al., 2003; Lee and Narayanan, 2005). For example, the repetition of the same dialogue act is likely to indicate trouble in communication.

2. Corpus Description

To date 29 experiments with 17 subjects have been carried out and transcribed, and we have collected 11.7 hrs of audio-visual data. Subjects ranged in age from 9 to 12. Some student subjects spent the entire tutoring session playing silently with the Legos, and declined to respond even when the tutor tried to engage them in conversation. However, some students were relatively loquacious, and provided us with a relatively rich speech database for our study. We manually extracted some utterances from users' speech that could reflect the user's cognitive activity. The database did not include back channels and "neutral" utterances such as *ok, thanks, that's nice,* etc. In addition, some data had to be discarded because of Lego block noise, heavy breathing, etc. To date we have collected 714 student utterances, containing approximately 50 minutes of relatively clean speech. On average each utterance has 4.2*s* speech and 8.1 words.

Three annotators worked on the Wizard-of-Oz audio data independently of each other. The annotation was performed based on speech content and dialogue context. The consistency among the annotators was annotated by the Kappa statistic, $K=(P_O-P_C)/(1-P_C)$, where P_O is the percent of times when the annotators agree, and P_C is the percent of times when the annotators are expected to agree by chance (Flammia, 1998). Kappa is 1.0 when the annotators agree on all transcriptions, and is 0.0 when the rate of agreement

equals the average rate that would be achieved by chance. The kappa statistics on our corpus annotation yielded a score of 0.93, indicating a very good agreement. Comparing with the inter-speaker agreement on emotion labeling achieved by Forbes-Riley and Litman (2004), Ang et al., (2002), etc., our annotator agreement is very high. This might because according to the annotation criteria, the distinction among the three cognitive states is more explicit than the distinction among the delicate and subtle emotional states. Therefore, the annotators are easy to make agreement on labeling. We used majority voting to resolve the annotation differences among the three annotators: for each utterance, if two or more annotators had the same label, then we assigned that label to the utterance, otherwise (i.e., each annotator had a different label) the utterance was removed from the corpus. Table 2 shows the distribution of utterances in our corpus with respect to the cognitive state classes.

4. Proposed approach

4.1 Overview of the Cognitive State Classifier

Our cognitive state classifier is based on the analysis of a large number of speech-related and language-related information sources, including long-term prosody, short-term spectrum, lexical features (word choice), and syntactic features (part-of-speech sequence). First, prosody conveys paralinguistic information about cognitive activities. For example, hesitant speech tends to have lower signal energy and longer phonetic word boundaries, while puzzled speech is often associated with raised pitch in the utterance-final syllable, especially but not exclusively if the utterance is a yes/no question. Second, some utterances are inherently ambiguous for cognitive state detection by means of prosody alone. For example, *wh*-questions and no-opinion statements (e.g., 'I don't know' denotes *puzzlement*, and 'I'm not sure' denotes *hesitation*) may be uttered with prosody indistinguishable from the prosody of a confident utterance. Spotting some key words/phrases embedded in the fluent speech helps to identify these special cases. Third, the cognitive state affects the articulation of phonemes, thus the spectra of speech signals may signal cognitive state. Fourth, we notice that speakers in confident mode and puzzled mode usually use different linguistic structures, while hesitant utterances are often

ungrammatical and incomplete. For example, a *wh*-question (usually indicates *puzzlement*) has different syntactic structure than a statement (often indicates *confidence*).

The overall structure of the cognitive state classifier is depicted in Figure 3. An automatic speech recognition (ASR) system first transcribes the utterance; word error rate of the ASR is typically high. Features for cognitive state classification are extracted as follows. First, lexical features are extracted from the recognized word string output of the ASR system. Second, the recognized word string is automatically tagged with part-of-speech, and the part-of-speech sequence is extracted as syntactic features. Third, the speech waveform and the transcribed/recognized word duration are used by the prosody analyzer: some prosodic features depend only on the speech waveform, but some also depend on word boundary alignment duration. Fourth, spectral evidence is obtained by passing the speech signal through three cognitive state-dependent ASR systems. The cognitive state-dependent ASR system is the same as the routine ASR system except that the former uses phone models that are adapted to data categorized into cognitive classes. The corresponding acoustic *log* likelihood scores based on different cognitive state assumptions are used as the spectral evidence. Finally all the analytical results go into a classification decision tree, which determines the cognitive state that maximizes the likelihood of the given acoustic, lexical, prosodic, and syntactic evidences.

4.2 Children Speech Recognition

4.2.1 Acoustic Analysis of Children Speech

Children's speech has very different acoustic characteristics than adults' speech. The vocal tracts of children are short and still growing. The shorter vocal tract length makes formant frequencies of children higher than those of adults. Speech of children under 13 years old exhibits both high inter-speaker and high intra- speaker variabilities, leading to extremely high speech-to-text word error rate (Narayanan and Potamianos, 2002). The *inter-speaker variability* includes age-dependent variability of formant frequencies, and unusually high variability among speakers of the same age caused by factors including the different growth rates of different children. The *intra-speaker variability* is caused by both developmental factors (children are less likely than adults to repeat an utterance with exactly the same articulation, so the spectral envelope shape both within a token and

across two repetitions has larger variations) and signal processing factors (the high *F*0 of a child means that formant peaks are not well represented in the spectrum, thus there is ambiguity in the mapping from cepstra to phones). These acoustic characteristics of children's speech result in degradation of ASR performance on children's speech. It has been reported that the in-vocabulary word error rate of children's speech is almost twice that of adults' speech (Zue et al., 2000).

We apply frequency warping (Lee and Rose, 1998; Zhan and Westphal, 1997) to normalize the vocal tract length of children, as shown in Figure 4. The frequency f is warped by means of a bilinear transform, which maps an input to an equal length of output in the frequency domain:

$$\varphi_{\beta_f}(f) = f + \frac{2 f_N}{\pi} \tan^{-1}\left(\frac{(1-\beta_f)\sin(\frac{f}{f_N}\pi)}{1-(1-\beta_f)\cos(\frac{f}{f_N}\pi)}\right),$$

where f_N is the Nyquist frequency, and the warping factor β_f is not constant but dependent on frequency. To compensate for the inter-speaker and intra-speaker variabilities, different warping factors are used on groups of children with the same age and gender. With reference to the published data (Lee et al., 1999), for each group, the warping factors at formants F_1, F_2 and F_3 are computed as the ratio of average formant values of that group to those of adult males. The warping factors for frequencies other than the three formants are approximated by piecewise linear interpolation.

4.2.2 Language Modeling

The construction of a bigram language model usually requires millions of words; the data available to us from the ITS Wizard-of-Oz simulations is far less than that amount. To make up for data sparsity, we derive from the Switchboard transcription (Godfrey et al., 1991) the POS-level linguistic information whereby to establish the word-level language model for our system. Our approach first promotes the words in our system vocabulary onto their POSs. A word often has several POSs. We consider all the possible POSs for a given word, and the probability of attaching a POS to that word is derived from the POS-tagged Switchboard transcription. Next we compute the POS-level bigram probabilities

based on the Switchboard transcription using the backoff smoothing technique (Katz, 1987). The POS tag of the transcription data is determined by an automatic tagger (Munoz et al., 1999). Then we apply the POS-level linguistic information to the system to derive a word-level bigram. Specifically, the word bigram model is given by:

$$p(w_2 \mid w_1) = \sum_{POS_2, POS_1} \{p(POS_1 \mid w_1) p((POS_2 \mid POS_1) p(w_2 \mid POS_2)\}$$

where w_1 and w_2 are two words in the vocabulary of our dialogue system, POS_1 and POS_2 are all possible part-of-speech tags of words w_1 and w_2, respectively, $p(POS_1 \mid w_1)$ is the empirical frequency with which POS_1 is attached to w_1 in the Switchboard transcription, $p(POS_2 \mid POS_1)$ is the POS-level bigram probability in the Switchboard transcription, and $p(w_2 \mid POS_2)$ is the empirical frequency of w_2 given POS_2 in the ITS Wizard-of-Oz speech data.

4.3 Lexical Analysis

Some keywords are closely associated with particular cognitive states. For example, filled pauses (e.g., 'uhm' and 'ahh') are likely to denote *hesitation*; wh-word such as what and when are likely to denote *puzzlement*. Data sparsity makes it impractical for us to automatically generate a keyword list. Therefore, we manually generate a list of approximately 50 key words/phrases that are affectively salient. In order to further avoid problems of data sparsity, the 50 key words/phrases are grouped into 9 keyword classes. The spotted key word(s)/phrase(s) from a recognized word string is (are) assigned to corresponding keyword classes, resulting in a 9-dimensional vector of binary indicator variables denoting the presence or absence of each keyword class. Table 3 lists all of the keyword classes and their keyword components.

4.4 Prosodic Analysis

As summarized in Table 4, the prosodic feature set includes:

Pitch: Pitch and probability of voicing are estimated using the FORMANT program in Entropic XWAVES. Pitch measurements are ignored in all frames with a low

probability of voicing. Valid pitch measurements are further standardized by the pitch range of the utterance:

$$pitch_f = \frac{pitch_f - \min_f(pitch_f)}{\max_f(pitch_f) - \min_f(pitch_f)}$$

where pitch_f is the pitch of the f^{th} frame, $\min_f(pitch_f)$ is the minimum non-zero pitch value of the entire utterance, and $\max_f(pitch_f)$ is the maximum pitch value of the entire utterance. We compare pitch in the end region (the final 100 ms) and the penultimate region (the previous 100 ms), as well as looking at the least-squares regression lines covering the two regions.

Energy-related features: Typically, an utterance falls to lower energy when close to completion. When an utterance ends in a sentence fragment, this fall has not yet occurred and thus energy remains high (Jurafsky et al., 1997). Therefore, the comparison of energy in the final and penultimate regions can be indicative of *hesitation*. Energy is measured on a log scale and then peak-normalized. The ratio of average energy in the final and penultimate regions of an utterance, and the first and second order time derivatives of energy in the final region are used as features.

Pause-related features: As a cue to *hesitation*, pause refers to a time period of non-speech lasting more than 600ms. Non-speech periods of less than 600ms are not called pause, here, because they are more likely to be minor disfluencies and intonational-phrase-final junctures, while longer pauses tend to mark perceptually prominent disfluencies and hesitation. To detect a pause, energy is first passed through an anti-symmetric edge-detection filter, and then is compared with a time-varying threshold (Li et al., 2002). Pauses of longer than 600 ms are extracted, their total duration is computed, and then normalized by the utterance duration.

Syllabic rate-related features: Usually people speak more slowly when they are hesitant and puzzled, as opposed to when they are confident. On average, the speaking rate of children is slower than that of adults. In addition, the speaking rate usually varies from person to person. Therefore, the syllabic rate of an utterance is divided by the speaker's normal syllabic rate, which is averaged from the speaker's utterances. In this study, syllabic rate is the average value of three estimators: (1) peak counting of the wide-band energy envelope followed by utterance duration

normalization; (2) pointwise correlation between pairs of subband energy envelopes, followed by peak counting and utterance duration normalization (Morgan and Fossler-Lussier, 1998); and (3) peak frequency of the energy modulation spectrum (Kitazawa et al., 1997).

Word-duration and utterance-duration features: Normal prosodic effects (prosodic phrase position and phrasal prominence) can lengthen a word by up to roughly 100%; duration increases of more than about 100% are likely to indicate *hesitation*. The duration feature in an utterance is characterized by a vector of four measurements: average and maximum of the normalized word durations (word durations are computed on the basis of alignment times produced by the ASR and normalized by the number of syllables) in the utterance, syllable-normalized utterance duration, and word-normalized utterance duration.

4.5 Spectral Analysis

Spectrum captures vocal-tract movements. Articulatory movements may be rapid or slow, extreme or reduced, depending on the level of arousal of the speaker; changes in the rate and extremity of articulation cause corresponding changes in the spectral correlates of phonemes. This hypothesis is verified by a study of Lieberman and Michaels (1962), which showed that human listeners can recognize emotional speech segments with an accuracy of 85%, but the accuracy decreased dramatically to just 47% when the spectral information was filtered out and only pitch and intensity information was preserved. Spectral feature is the least frequently employed feature in published studies of affect recognition. In previous studies the only case of incorporating spectral features into emotion recognition is the study of Polzin and Waibel (2000), who used an HMM-based speech recognition system to classify emotion. Polzin and Waibel modeled spectral information using a triphone Gaussian mixture model, in which each phone has 3 states, and each state observes 32 MFCCs. The study did not show a significant contribution of spectral information to emotion recognition.

In this study, we propose a new approach for incorporating spectral information into user affect recognition. The short-term spectral envelope and its first and second temporal derivatives capture the dynamics of phoneme articulation. To manifest the influence of cognitive states on phoneme articulation, we extract spectral features and

adapt phoneme models using utterances partitioned into the three cognitive state categories. For example, the *confidence*-dependent phoneme models are adapted using the subset of training utterances from the ITS corpus that have been labeled as *confidence*. For a given utterance, ASR based on each cognitive state-dependent model generates a word string associated with an acoustic likelihood score. Each acoustic likelihood score spans a wide range. We compare the acoustic likelihood scores based on different cognitive state assumptions, and use their differences as the spectral cues. The spectral features are listed in Table 5.

4.6 Syntactic Analysis

The syntactic composition of sentences is useful to reveal cognitive state. *Puzzlement* is often expressed in questions or in statements beginning with a first-person pronoun (*I don't understand*), while *confidence* is often expressed in statements about the task. Questions and statements can be distinguished, with reasonable accuracy, based on part-of-speech of the first three words of the utterance. Sentences expressing *hesitation* are often ungrammatical or incomplete. The final word, or the exit of a sentence, is especially useful in detecting the completeness or incompleteness of a sentence. For example, an exit word with a conjunction part-of-speech usually indicates that the sentence is incomplete. The context of the exit word usage is also important. For example,

This gear is a much bigger

This gear is much bigger

The first sentence is incomplete while the second sentence can be considered complete, although they have the same exit word.

Spoken language usually contains ungrammatical junctures and dysfluencies, and therefore a complete syntactic parse of a spoken utterance may be inefficient or impossible. Instead, we use the part-of-speech of the first three words and the last three words of a sentence as the syntactic cues for cognitive state classification. These six words, as shown in our corpus analysis, rarely contain dysfluencies; information about these six words is rarely sufficient to reconstruct a complete parse of the sentence, but the correct part-of-speech of these six words is often sufficient to correctly classify cognitive state of the user. Part-of-speech tagging is performed by an automatic tagger (Munoz et al., 1999).

5. System Evaluation

5.1 Children's Speech Recognition

We designed a triphone-based ASR system with the help of the HTK toolkit. Sampled at 11 KHz, the speech input was pre-emphasized and grouped into frames of 330 samples with a window shift of 110 samples. The speech signal was characterized by 13 Mel-frequency cepstral coefficients (MFCCs) normalized by cepstral mean subtraction and log-scaled energy normalized by the peak. MFCCs, energy, their deltas and delta-deltas together formed a vector of 42 features. In the system, each word was represented by the concatenation of the models of its component phones. Each phone model was a left-to-right 3-state HMM with an output distribution of 16 Gaussian mixtures per state. The speaker-independent universal background phoneme model was trained from the TIMIT database. Recognition was accomplished by a frame synchronous token-passing search algorithm to determine the sequence of words that maximized the likelihood of a given utterance.

Because of the limited size of the database, the 17 child speakers were divided into 6 groups in terms of age and gender (see Table 6). Children's speech has high inter-speaker variability, which means that pitch and formants are dependent on age. The age-dependent formants introduce variability in spectral features across age groups, so if a phoneme model for ASR is based on a certain age group and tested against another age group, the mismatch between training and test data will result in degradation. According to the study of Lee et al. (1999), pitch and formant frequencies of male and female talkers become distinct at age 11. Therefore, gender is not a factor causing the degradation in speech recognition for children 9-10 years old. The universal background model trained from TIMIT was further adapted to the groups of children (using about half of the total speech of each group) based on maximum likelihood linear regression adaptation followed by maximum *a posteriori* adaptation. The recognition performance is listed in Table 6.

The low recognition accuracy on the one hand was caused by the acoustic characteristics of children's speech as we have described, and on the other hand was caused by the linguistic correlates of children's speech. Children's speech has higher

degree of spontaneity. Children's speech is prone to dysfluency containing mispronunciation, false starts, breath noise, and filled pauses, more obvious with children younger than 11 years old (Potamianos and Narayanan, 2001). Since the children users in our system were not familiar with the experiment contents, their utterances were even more incoherent and dysfluent than those of typical dialogue system users, e.g., *Ahmm when you… after it goes around once, the other one goes around the same, the same, I mean it goes around… you know you only have to spin it around once, and that makes sense basically because they are the same size.* The language characteristics of children's speech in our system caused the language model perplexity to be as high as 593.42.

5.2 Cognitive State Classification Performance

The information sources were integrated probabilistically via a classification decision tree. We chose to use a decision tree classifier because of its efficiency in handling samples with high dimensionality, mixed data types, and nonstandard data structure. For the decision tree program, we used *See5*, which is a data mining tool for discovering patterns or relationships in data, assembling them into classifiers that are expressed as decision trees or sets of *if-then* rules, and using them to make valid predictions (Rulequest Research, 2004). Each variable in the feature vector $\mathbf{x} = (x_1, x_2 \ldots x_n)$ was in continuous, discrete, or character type. There were a total of 33 features, as listed in Table 7. The See5 program was invoked with the options *rulesets* (meaning that tree-based classifiers were backed off into collections of *if-then* rules), *boost* (See5 generates several classifiers, whose classifications are weighted and summed to determine the predicted class), *ignore costs file* (so that the training algorithm gave equal weight to false positive and false negative errors), and *global pruning* (a large tree is first grown to fit the data closely and is then pruned by removing parts that are predicted to have a relatively high error rate).

We randomly chose 400 (56%) samples for training and the remaining 314 (44%) samples for evaluation; this process was repeated 10 times, and the results averaged (10-fold cross-validation). Our study showed that for syntactic analysis, counting only the last word (exit word) yielded slightly better results than processing the last three words in an utterance, thus we recorded the result in the former case.

5.2.1 Transcribed Speech

We first tested cognitive state classification using manual transcriptions of the Wizard-of-Oz corpus. The spectral features and duration-based features were derived from the forced-alignment of the manual transcription to the waveform. We computed precision, recall, and *F*-score ($f = \frac{1}{0.5/p + 0.5/r}$) for each of the three cognitive states, and presented the results in Table 8. We further presented in Figure 5 the *F*-score values of the three cognitive state classifications using different features. The figure shows that when spectrum is excluded for accounting, the best classification performance falls on *confidence,* followed by *puzzlement*. However, there is no distinction among the three state classifications when spectrum is used.

In Figure 6, we compare the efficiency of individual features in terms of classification accuracy and the *F*-score-based evaluation measures. Classification accuracy is the fraction of the test set that is correctly classified with respect to the three classes. *F*-score is a harmonic mean of precision and recall: *F*-score is high only when both precision and recall are high. Therefore, *F*-score is a more reliable measure. However, the *F*-score values that we have derived were class-dependent as shown in Figure 5. We were interested in the global performance measure across the classes, so we designed two global *F*-score measures: 1) averaged *F*-score over the three state classifications; and 2) $f_{ave} = \frac{1}{0.5/p_{ave} + 0.5/r_{ave}}$, where p_{ave} was the averaged precision over the three state classifications, and r_{ave} was the averaged recall over the three state classifications. The study results showed that spectrum greatly outperformed the other features in cognitive state classification. When all of the variables were combined, the classification accuracy was slightly increased (2.1% absolute) relative to a classifier using spectrum only. Part-of-speech played the least significant role in the classification, partly because of errors existing in the automatic part-of-speech tagger. The relatively low accuracy using word-related variables (lexicon and part-of-speech) demonstrated that the cognitive state variables were associated with speech features much more than text features. We also compared the prosodic features in terms of their occurrence frequency

in the decision tree rules. Among the several features available to the classifier, the energy-based features were used most often, followed by the duration-based features.

5.2.2 Recognized Speech

Our system performance using recognized speech is listed in Table 9, and the *F*-score values of the three state classifications based on different features are shown in Figure 7. The classification results showed that the recognition of *confidence* outperformed the recognition of *puzzlement* and *hesitation* in terms of *F*-score value. The syntactic features failed to converge during machine training, and thus were not be used for classification. We compared the results yielded by the prosodic, lexical, and spectral features. Same as transcribed speech, the test results on recognized speech showed that spectrum was the exceptionally useful variable for cognitive state classification. For recognized speech, the combination of all feature variables yielded the highest accuracy 95.7%. The ranking of the different feature vectors (excluding the syntactic feature) according to classification accuracy and global *F*-score measures are shown in Figure 8. Similar to the conclusion drawn from Figure 6, spectrum played the most important role, followed by prosody. Moreover, the combination of spectrum and prosody almost decided the classification performance, no matter whether lexicon and part-of-speech were included or not. The comparison of prosodic features in terms of occurrence frequency in the decision tree rule set showed that duration is the most important prosodic feature, followed by energy. In contrast to many other studies (Mozziconacci and Hermes, 1998; Juang and Furui, 2000; Petrushin, 2000; Kang et al., 2000), pitch played a non-significant role in the detection of cognitive states, probably because of the frequent occurrence of pitch tracking inaccuracies in our noisy speech data. It was hard to discriminate voiced regions from unvoiced regions in the noisy recording environment. The energy of unvoiced regions carried information irrelevant to the pitch estimate, and thus the automatically extracted pitch might have contained too many pitch tracking errors to be an efficient feature.

5.2.3 Comparison of Transcribed Speech and Recognized Speech

We compared the classification results using automatically recognized speech with those using manually transcribed speech to demonstrate the robustness of cognitive state

classification to speech recognition errors. We list the classification accuracy on transcribed speech, recognized speech, and the relative reduction in Table 10. The test results suggested that spectrum and prosody were robust to speech recognition errors. The prosody-based classification accuracy was even slightly better (by 1.79% absolute) using recognized speech rather than transcribed speech, possibly because the forced-alignment procedure did not guarantee accurate estimates of word durations. Spectrum-based classification accuracy was a bit worse using recognized speech (4.65% absolute). When spectrum and prosody were combined, cognitive state classification accuracy was almost identical for manual and automatic transcriptions. The keyword class-based classification was considerably worse using recognized speech (18.4% absolute), demonstrating the sensitivity of keyword features to the speech recognition errors.

6. Conclusion

This paper has proposed a three-way cognitive state classification designed to be useful for an intelligent tutorial application. We proposed classifying the cognitive activities of children users during their learning process into three categories: *confidence*, *puzzlement* and *hesitation*. The task was performed on 714 spontaneous utterances extracted from the audio-visual data that was collected in a Wizard-of-Oz simulation of the tutorial system. This particular cognitive state classification task could achieve an inter-annotator agreement of kappa score 0.93, higher than emotion labelling in dialogue systems. Automatic cognitive state classification used features including keyword classes, prosody, spectrum, and syntax. Test results showed that the classification accuracy of the cognitive state classifier reached up to 96.6% for manually transcribed speech and 95.7% for automatically recognized speech, indicating that cognitive state classification is both accurate and robust to speech recognition errors. In particular, the test results showed that the proposed spectral features, measured as the difference in acoustic log likelihood between different cognitive-state dependent speech recognizers, much outperformed the other features, providing exceptionally accurate estimate of the cognitive state. Moreover, the test results showed that prosodic features and spectral features were both robust to speech recognition errors, much more than the lexical and syntactic features. In addition, among the prosodic features, the duration-based features and energy-based features were

found to participate in the *if-then* decision more frequently than the other features; the pitch-based features played the least important role, possibly because of pitch tracking errors caused by the noisy recording environment.

This paper has described the task of cognitive state detection as an end in itself, but it is our goal and intention to use cognitive state detection as part of a larger automatic speech understanding system for the intelligent tutorial application. In considering possible applications of this work, we propose that an automatic labelling scheme that can be computed with an accuracy of 95.7% would probably be a useful input feature for further stages of automatic speech understanding and dialogue control.

Acknowledgement

We would like to thank Brian Pianfetti for providing us with the ITS Wizard-of-Oz audio data, and Thomas S. Huang for providing us with all kinds of help in our work. This work is supported by NSF grant number 0085980. Statements in this paper reflect the opinions and conclusions of the authors, and are not endorsed by the NSF.

REFERENCES

Alpert, S. R., Singley, M. K., and Carroll, J. M. 1999. Multiple instructional agents in an intelligent tutoring system. Intl. Workshop on Instructional Uses of Animated and Personified Agents (at the 9th Intl. Conf. on AI in Education).

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. Proc. of ICSLP.

Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E. 2003. How to find trouble in communication. Speech Communication. 40, 117-143.

Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J., and Fischer, K. 2000. The recognition of emotion. In W. Wahlster (ed.) Verbmobil: Foundations of Speech-to-Speech Translations. Springer, New York, Berlin.

Beck, J., Jia, P., and Mostow, J. 2003. Assessing student proficiency in a reading tutor that Listens. Proc. of the 9th Intl. Conf. on User Modelling.

Clark, B., Fry, J., Ginzton, M., Peters, S., Pon-Barry, H., Thomsen-Gray, Z. 2001. A multimodal intelligent tutoring system for shipboard damage control. Proc. of 2001 Intl. Workshop on Information Presentation and Multimodal Dialogue. Verona, Italy.

Cole, R., van Vuuren, S., Pellom, B., Hacioglu, K., Ma, J., Movellan, J., Schwartz, S., Wade-Stein, D., Ward, W., Yan, J. 2003. Perceptive animated interfaces: first steps toward a new paradigm for human-computer interaction. Proceedings of the IEEE: Special Issue on Multimodal Human Computer Interface.

Corbett, A. T. and Anderson, J. R. 1992. The Lisp intelligent tutoring system: research in skill acquisition. In J. Larkin and R. Chabay (eds.) Computer Assisted Instruction and Intelligent Tutoring System: Shared Goals and Complementary Approaches. Lawrence Eflbaum, Hillsdale, NJ.

Crowley, R., Medvedeva, O., and Jukic, D. 2003. SlideTutor: a model-tracing intelligent tutoring system for teaching microscopic diagnosis. Proc. of the 11th Intl. Conf. on Artificial Intelligence in Education. Sydney, Australia.

Fernandez, R. and Picard, R. W. 2003. Modeling driver's speech under stress. Speech Communication, 40, 145-159.

Flammia, G. 1998. Discourse segmentation of spoken dialogue: an empirical approach. Ph.D. Thesis. MIT.

Forbes-Riley, K. and Litman, D. 2004. Predicting emotion in spoken dialogue from multiple knowledge sources. Proc. of HLT/NAACL.

Gobl, C. and Chasaide, A. N. 2003. The role of voice quality in communicating emotion, mood and attitude. Speech Communication, 40, 189-212.

Godfrey, J.J., Holliman, E.C., and McDaniel, J., 1991. SWITCHBOARD: Telephone speech corpus for research and development. Proc. ICASSP, 517-520.

Graesser, A. C., Lehn, K. V., Rose, C. P., Jordan, P. W., and Harter, D. 2001. Intelligent tutoring system with conversational dialogue. AI Magazine. *22*(4), 39-52.

Juang, B.-H. and Furui, S. 2000. Automatic recognition and understanding of spoken language—a first step towards natural human-machine communication. *Proc. IEEE* 88(8): 1142-1165.

Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., and Ess-Dykema, C. V. 1997. Switchboard discourse language modeling project final report. Johns Hopkins LVCSR Workshop.

Kafai, Y., and Harel, I. 1991. Children learning through consulting: when mathematical ideas, knowledge of programming and design, and playful discourse are intertwined. In: Harel I. and Papert S. (Ed.), Constructionism, Ablex, Norwood, NJ, pp. 111-140.

Kang, B.-S., Han, C.-H., Lee, S.-T., Youn, D.-H., Lee, C. 2000. Speaker dependent emotion recognition using speech signals. Proc. ICSLP.

Katz, S. M. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Trans. Acoustics, Speech and Signal Processing. 35(3), 400-401.

Kitazawa, S., Ichikawa, H., Kobayashi, S., and Nishinuma, Y. 1997. Extraction and representation rhythmic components of spontaneous speech. Proc. of EuroSpeech. Rhodes, Greece.

Lee, C. M. and Narayanan, S. 2005. Toward detecting emotions in spoken dialogs. IEEE Trans. on Speech and Audio Processing, 13(2): 293-303.

Lee, L. and Rose, R. 1998. A frequency warping approach to speaker normalization. IEEE Trans. on Speech and Audio Processing, 6(1): 49-59.

Lee, S., Potamianos, A., and Narayanan, S. 1999. Acoustics of children's speech: developmental changes of temporal and spectral parameters. The Journal of the Acoustical Society of America. 105(3), 1455-1468.

Lesgold, A., Lajoie, S., Bunzo, M., and Eggan, G. 1990. Sherlock: a coached practice environment for an electronics troubleshooting job. In J. Larkin, R. Chabay, and C. Sheftic (eds.) Computer Assisted Instruction and Intelligent Tutoring Systems Establishing Communication and Collaboration. Erlbaum, Hillsdale, NJ.

Li, Q., Zheng, J., Tsai, A., and Zhou, Q. 2002. Robust endpoint detection and energy normalization for real-time speech and speaker recognition. IEEE Transactions on Speech and Audio Processing. 10(3), 146-157.

Lieberman, P. and Michaels, S. B. 1962. Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. Journal of the Acoustical Society of America, 34: 922-927.

Litman, D., Rose, C., Forbes-Riley, K., VanLehn, K., Bhembe, D., and Silliman, S. 2004. Spoken versus typed human and computer dialogue tutoring. Proc. of the 7th Intl. Conf. on Intelligent Tutoring Systems, Maceió, Brazil.

Litman, D. and Silliman, S. 2004. ITSPOKE: an intelligent tutoring spoken dialogue system. Proc. of the HLT/NAACL. Boston, MA.

Martinovsky, B. and Traum, D. 2003. Breakdown in human-machine interaction: the error is the cue. Proc. of the ISCA Tutorial and Research Workshop on Error Handling in Dialogue Systems.

Morgan, N. and Fosler-Lussier, E. 1998. Combining multiple estimators of speaking rate. Proc. of ICASSP. Seattle, WA.

Mostow, J., Beck, J., Winter, S. V., Wang, S., and Tobin, B. 2002. Predicting oral reading miscues. Proc. of ICSLP.

Mozziconacci, S. and Hermes, D. 1998. Study of intonation patterns in speech expressing emotion or attitude: production and perception. IPO Annual Progress Report, IPO, Eindhoven.

Munoz, M., Punyakanok, V. Roth, D., and Zimak, D. 1999. A learning approach to shallow parsing. In EMNLP-WVLC '99.

Narayanan, S., and Potamianos, A. 2002. Creating conversational interfaces for children. IEEE Transactions on Speech and Audio Processing. 10(2), 65-78.

Pellom, B., Ward, W., and Pradhan, S. 2000. The CU communicator: an architecture for dialogue systems. Proc. of ICSLP, Beijing, China.

Petrushin, V. 1999. Emotion in speech: recognition and application to call centers.

Petrushin, V. A. 2000. Emotion recognition in speech signal: experimental study, development, and application. Proc. ICSLP.

Polzin, T. S. and Waibel, A. 2000. Emotion-sensitive human-computer interfaces. ISCA Workshop on Speech and Emotion: a Computational Framework for Research.

Pon-Berry, H., Clark, B., Bratt, E. O., Schultz, K. and Peters, S. 2004. Evaluating the effectiveness of SCoT: a spoken conversational tutor. Proceedings of ITS 2004 Workshop on Dialogue-based Intelligent Tutoring Systems.

Potamianos, A. and Narayanan, S. 2001. Robust recognition of children's speech. IEEE Trans. Speech and Audio Processing.

Reyes, R. L., Galvey, C., Gocolay, M. C., Ordona, E., and Ruiz, C. 2000. Multimedia intelligent tutoring system for context-free grammar. Proc. Philippine Computing Science Congress.

Rulequest Research. 2004. Data Mining Tools. http://www.rulequest.com/see5-info.html

Schultz, K., Bratt, E. O., Clark, B., Peters, S., Pon-Barry, H., Treeratpituk, P. 2003. A scalable, reusable spoken conversational tutor: SCoT. AIED 2003 Supplementary Proceedings.

Steele, M. M. and Steele, J. W. 1999. Discover: an intelligent tutoring sytem for teaching students with learning difficulties for solve word problems. Journal of Computers in Mathematics and Science Teaching. 18(4), 351-359.

Ward, W. and Pellom, B. 1999. The CU communicator system, IEEE Workshop Automatic Speech Recognition and Understanding, Keystone, CO.

Wilensky, U. 1991. Abstract meditations on the concrete. In: Harel I. and Papert S. (Ed.), Constructionism, Ablex, Norwood, NJ.

Zhan, P. and Westphal, M. 1997. Speaker normalization based on frequency warping. Proc. of ICASSP.

Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., and Hetherington, L. Jupiter: a telephone-based conversational interface for weather information. IEEE Trans. Speech Audio Processing, 8: 85-96.

Table 1: An excerpt of the tutorial dialogue scenario

(1) T: What are you exploring there?

(2) U: Seeing if the small gears move the big gear.

(3) T: What can you tell me about the directions they are spinning?

(4) U: They're moving in different directions mostly.

(5) T: What else do you notice?

(6) U: Big gears move in different ways and uhm … with the first when you push one of the first gears, the other gear, the last gear moves, you know, and the gear after that moves then the gear after that one moves.

(7) T: What if we try just using three gears? What do you notice?

(8) U: They're moving in the same direction.

(9) T: What effect does the medium gear have?

(10) U: The medium gear is, it is stronger than the big one and it's smaller and weaker than the small one.

(11) T: Why is the smaller gear stronger?

(12) U: Their teeth are smaller. If you look at it very closely, then there's just a little bit more space.

Table 2: Cognitive state statistics

	Confidence	***Puzzlement***	***Hesitation***
# of Utterances	441	216	57
% of All Utterances	61.8	30.2	8.0

Table 3: Keyword classes and component keywords

Keyword Class	***Keywords***
Affirm	yes, yeah, yep, no
Digit	one, two, three, etc
Known	I know, I believe
Uhm	ahh, ahm, ahmm, uhm
Reason	because, so
Unknown	don't know, don't understand
Auxiliary	can I, can you, could you, do I, do you, should I, should we, would it, would you, would it
Uncertain	not sure, not exactly sure
Wh-word	how, what, when, where, which, why

Table 4: List of the prosodic features

Feature	***Description***
F0_ratio	Ratio of mean *F*0 over the end region (the final 100 ms) and the penultimate region (the previous 100 ms).
F0_reg_pen	Least-squares all-points regression over the penultimate region.
F0_reg_end	Least-square all-points regression over the end region.
F0_norm	The number of nonzero *F*0 frames normalized by the utterance duration.
loge_ratio	Ratio of logarithmic energy over the end region and the penultimate region.
derive_logE	Mean of peak-normalized logarithmic energy derivative over the end region.
acce_logE	Mean of peak-normalized logarithmic energy acceleration over the end region.
norm_pause	Total pause durations normalized by the utterance duration.
Syllarate	Syllabic rate normalized by the speaker's normal speaking tempo.
mean_norm_word_dur	Mean of word duration which is normalized by the number of syllables the word has.
max_norm_ word_dur	Maximum of word duration which is normalized by the number of syllables in that word.
utt_dur_by_syllable	Utterance duration normalized by the number of syllables in that utterance.
utt_dur_by_word	Utterance duration normalized by the number of words in that utterance.
max_abs_word_dur	Maximum of absolute word duration.

Table 5: List of the spectral features

Feature	***Description***
$Spect_c$-$Spect_p$	ASR log likelihood using the *confidence*-adapted phone models minus ASR log likelihood using the *puzzlement*-adapted phone models
$Spect_c$-$Spect_h$	ASR log likelihood using the *confidence*-adapted phone models minus ASR log likelihood using the *hesitation*-adapted phone models
$Spect_p$-$Spect_h$	ASR log likelihood using the *puzzlement*-adapted phone models minus ASR log likelihood using the *hesitation*-adapted phone models

Table 6: Speech recognition rate of children, by age group (vocabulary size = 515). Test is against the total minutes of speech, including the adaptation data.

Age	***Gender***	***Total Minutes (Adaptation Minutes)***	***Recognition Accuracy, %***
9	M & F	3.68 mins (1.8 mins)	34.45
10	M & F	10.14 mins (5 mins)	40.27
11	M	11.02 mins (5 mins)	61.58
	F	5.99 mins (3 mins)	50.13
12	M	13.61 mins (6mins)	60.43
	F	2.99 mins (1.5 mins)	54.20

Table 7: List of all features used for cognitive state classification

Feature	***Size***	***Description***
Prosody	14	Continuous: 4 pitch-related; 3 energy-related; pause; syllabic rate; 5 duration-related.
Lexicon	9	Binary: absence vs. presence of each keyword class.
Part-of-speech	6	Character: either a character such as *VBP*, or a blank in the case that the utterance is shorter than 6 words.
Spectrum	3	Continuous
Target	1	Discrete: 1=*confidence*, 2=*puzzlement*, 3=*hesitation*.

Table 8: Precision *p*, recall *r*, and *F*-score *f* using transcribed words. Pros = Prosody, Lex = Lexicon, Pos = Part-of-speech, Spect = Spectrum.

	Confidence			***Puzzlement***			***Hesitation***		
Feature	*p*	*r*	*f*	*p*	*r*	*f*	*p*	*r*	*f*
Pros	0.896	0.942	0.918	0.791	0.777	0.784	0.710	0.484	0.576
Lex	0.818	0.914	0.864	0.850	0.742	0.792	0.634	0.417	0.503
Pos	0.712	0.967	0.820	0.862	0.479	0.616	0.357	0.019	0.036
Spect	0.938	0.977	0.957	0.951	0.899	0.924	0.983	0.913	0.947
Pros+Spect	0.952	0.988	0.970	0.971	0.917	0.944	0.984	0.909	0.945
Pros+Lex+ Pos+Spect	0.960	0.987	0.974	0.971	0.939	0.955	0.988	0.919	0.953

Table 9: Precision *p*, recall *r*, and *F*-score *f* using recognized words. Pros = Prosody, Lex = Lexicon, Pos = Part-of-speech, Spect = Spectrum.

	Confidence			***Puzzlement***			***Hesitation***		
Feature	*p*	*r*	*f*	*p*	*r*	*f*	*p*	*r*	*f*
Pros	0.904	0.940	0.922	0.811	0.856	0.833	0.860	0.474	0.611
Lex	0.636	0.922	0.753	0.700	0.230	0.346	0.221	0.072	0.109
Spect	0.889	0.962	0.924	0.930	0.834	0.880	0.876	0.720	0.790
Pros+Spect	0.962	0.979	0.970	0.957	0.927	0.942	0.902	0.881	0.891
Pros+Lex+ Pos+Spect	0.961	0.977	0.969	0.963	0.922	0.942	0.913	0.930	0.921

Table 10: Comparison of classification correctness between transcribed speech and recognized speech. Pros = Prosody, Lex = Lexicon, Pos = Part-of-speech, Spect = Spectrum.

Feature	***Transcribed Speech***	***Recognized Speech***	***Recognized – Transcribed (% relative)***
Pros	85.4%	87.2%	2.1%
Lex	81.6%	63.2%	-22.6%
Pos	73.6%	-	-
Spect	94.6%	90.0%	- 4.8%
Pros+Spect	96.0%	95.5%	- 0.5%
Pros+Lex+Pos+Spect	96.6%	95.7%	-0.9%

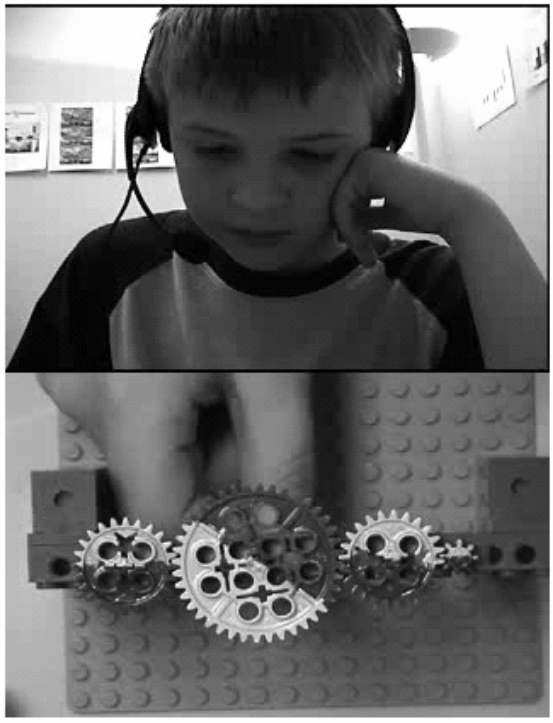

Figure 1: Graphic demonstration of the lego playing scene (the lower picture is showing the Lego gears that the child in the upper picture is playing).

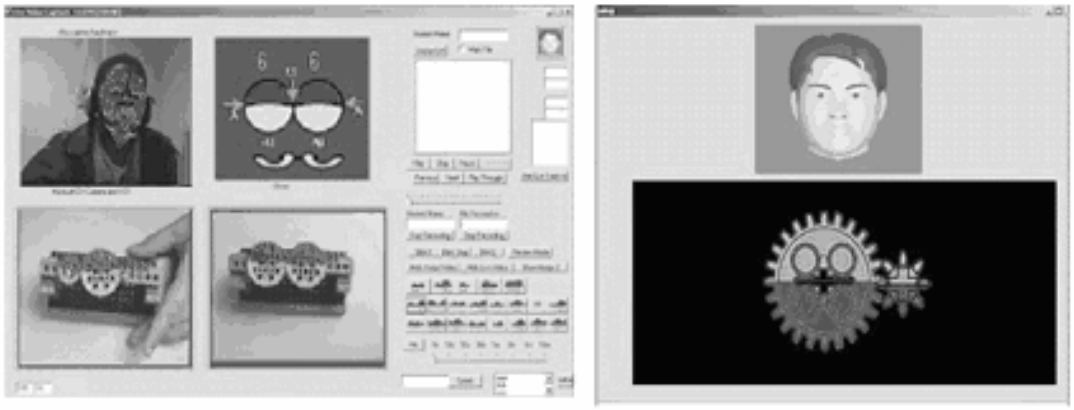

Figure 2: Screen displays on the tutor's computer (left) and the user's computer (right)

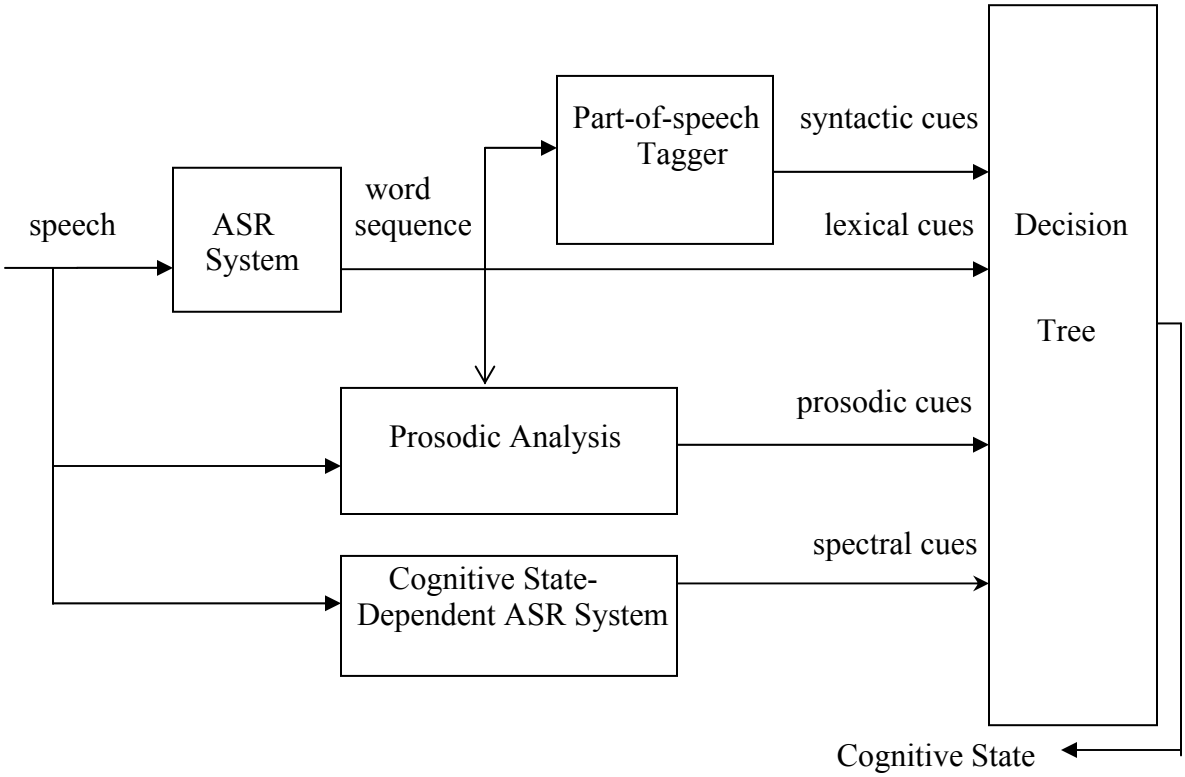

Figure 3: Architecture of the cognitive state classification system

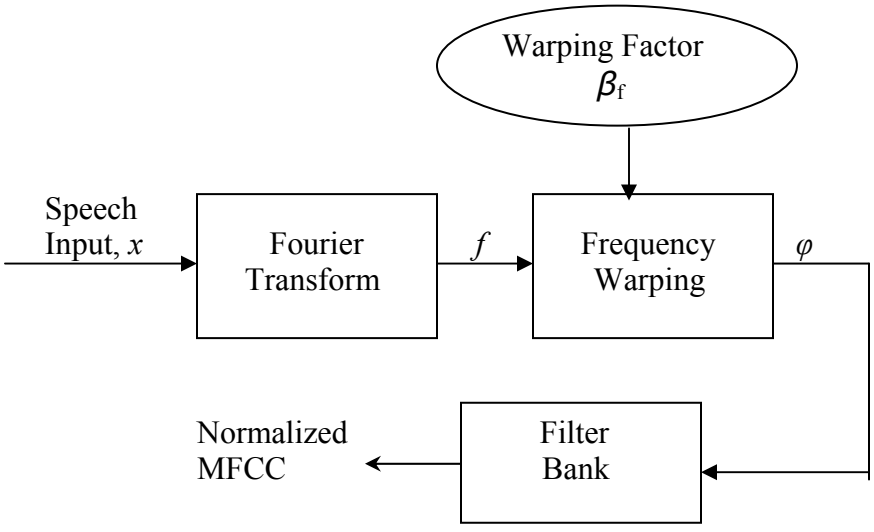

Figure 4: Vocal tract length normalization by frequency warping

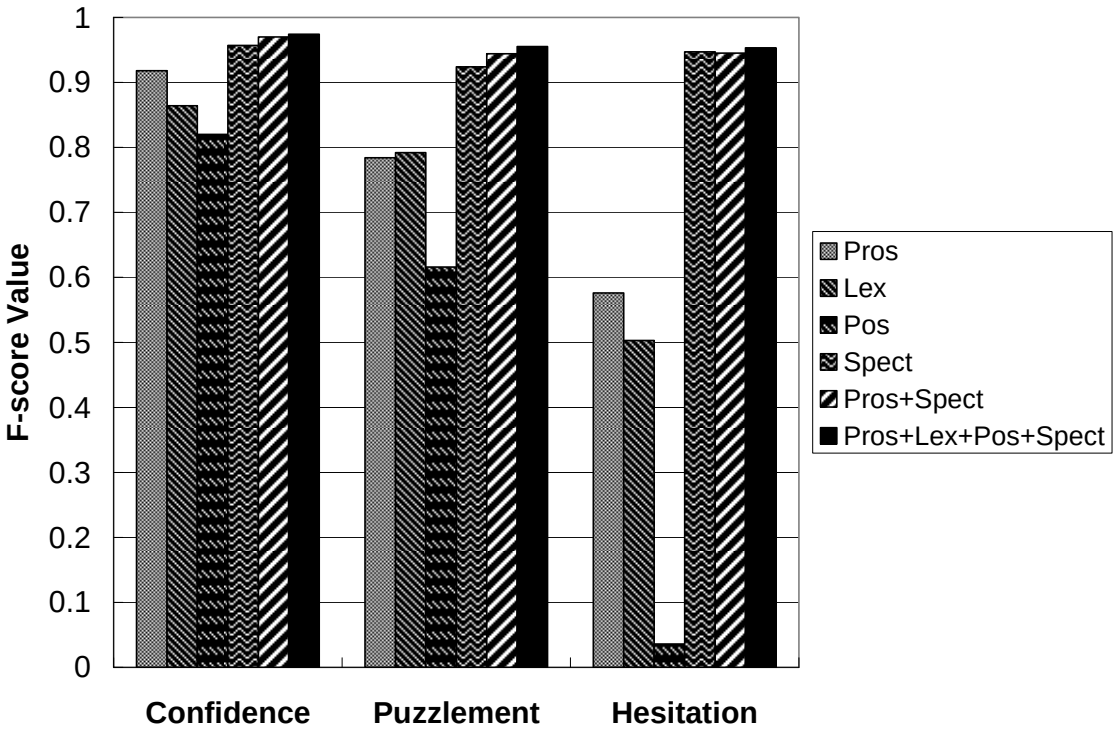

Figure 5: Classification performance with transcribed speech. Pros = Prosody, Lex = Lexicon, Spect = Spectrum, Pos = Part-of-speech.

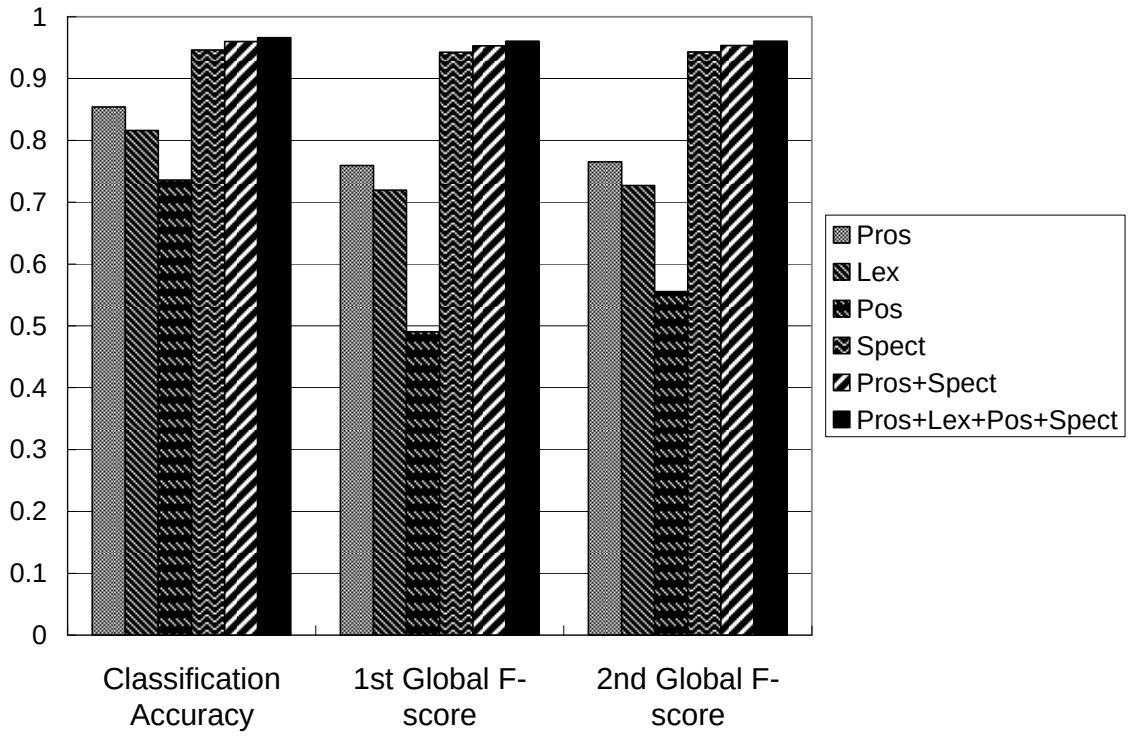

Figure 6: Feature accuracy ranking in classification of transcribed speech, in terms of classification accuracy and two global *F*-score measures: the first measure is the averaged three one-class *F*-scores; the second measure is *F*-score computed from the average precision and average recall of the three one-class classifications. Pros = Prosody, Lex = Lexicon, Pos = Part-of-speech, Spect = Spectrum.

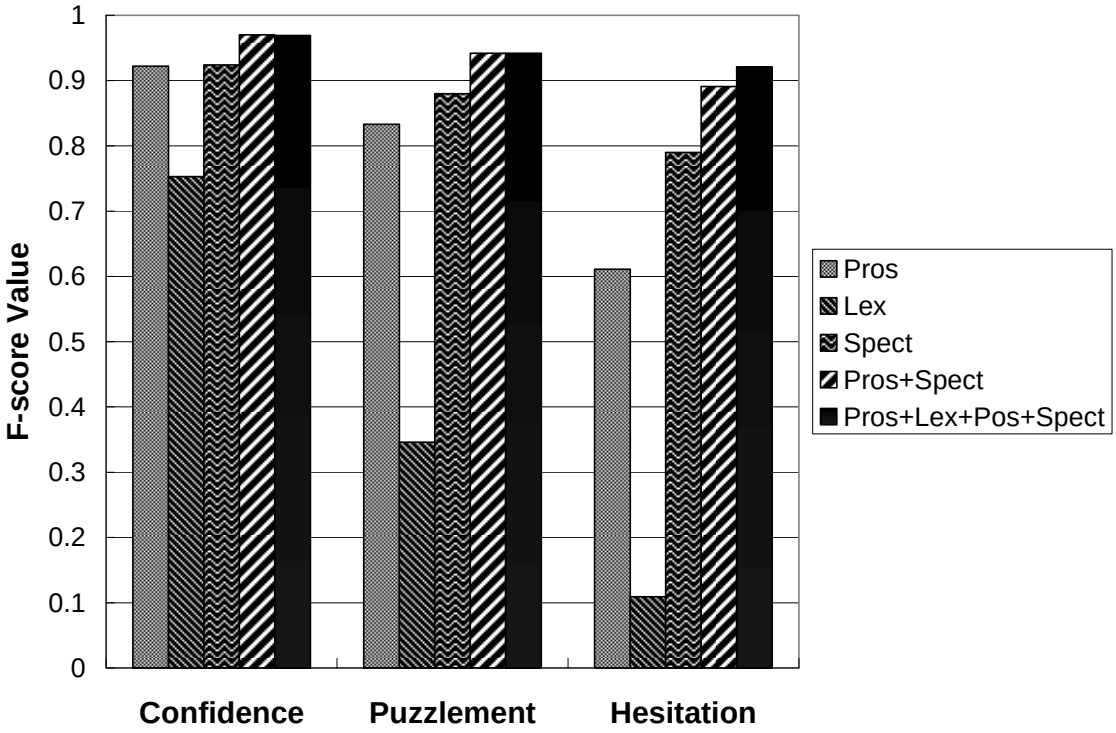

Figure 7: Classification performance with recognized speech. Pros = Prosody, Lex = Lexicon, Spect = Spectrum, Pos = Part-of-speech.

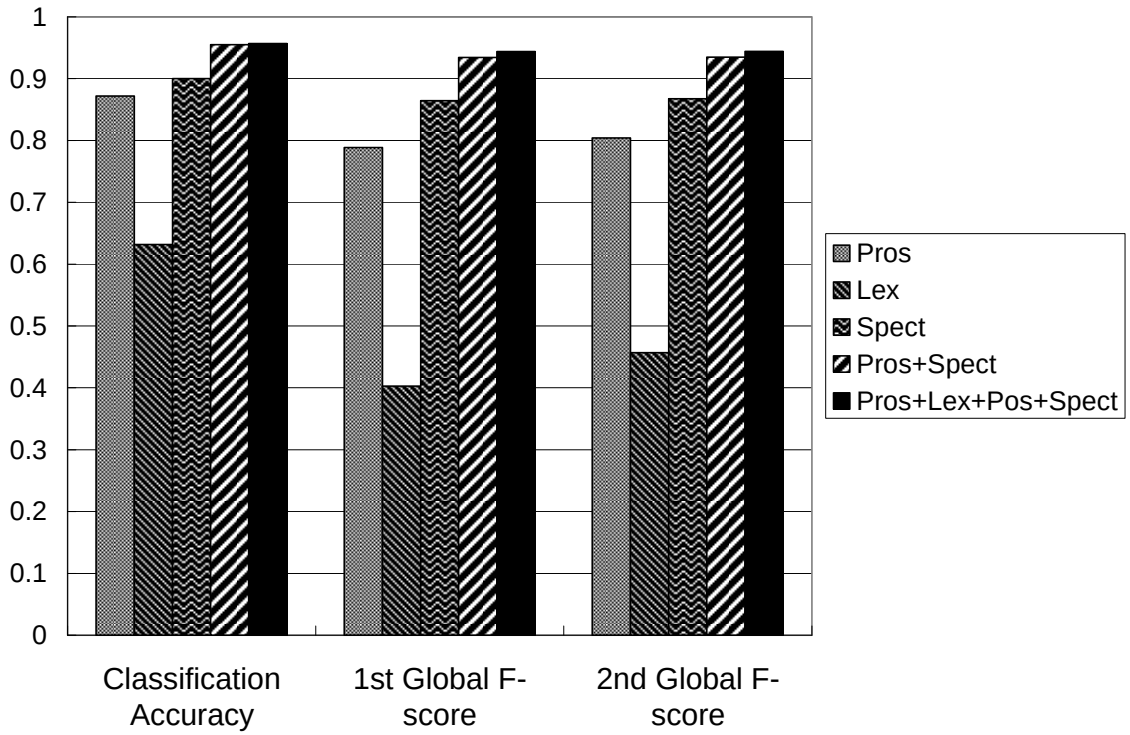

Figure 8: Feature accuracy ranking in classification of recognized speech, according to the classification accuracy and two global *F*-score measures: the first measure is the averaged three one-class *F*-scores; the second measure is *F*-score computed from the average precision and average recall of the three one-class classifications. Pros = Prosody, Lex = Lexicon, Spect = Spectrum, Pos = Part-of-speech.