ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13-16, 2003

MENTAL STATE DETECTION OF DIALOGUE SYSTEM USERS VIA SPOKEN LANGUAGE

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

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
{tzhang1,hasegawa,sel}@ifp.uiuc.edu

ABSTRACT

This paper presents an approach to simulate the mental activities of children during their interaction with computers through their spoken language. The mental activities are categorized into three states: confidence, confusion and frustration. Two knowledge sources are used in the detection. One is prosody, which indicates utterance type and user's attitude. The other is embedded key words/phrases which help interpret the utterances. Moreover, it is found that children's speech exhibits very different acoustic characteristics from adults. Given the uniqueness of children's speech, this paper applies a vocal-tract-length-normalization (VTLN)-based technique to compensate for both inter-speaker variability and intra-speaker variability in children's speech. The detected key words/phrases are then integrated with prosodic information as the cues for the MAP decision of mental states. Tests on a set of 50 utterances collected from the project experiment showed the classification accuracy was 74%.

1. INTRODUCTION

Our dialogue system is designed for an NSF-funding project, titled "Multimodal Human Computer Interaction System: Toward a Proactive Computer." The testbed of this project is an environment for education in science and technology, using the Lego-Logo Mindstorms construction set, with children of primary school age. The emphasis is on developing a proactive computer agent to encourage the interests of kids in science and technology. The communication between the computer agent and the user is via a human-computer dialogue system. The system design is based on two assumptions: 1) the proactive assumption: the computer can initiate communications when the user needs guidance, assistance, and encouragement rather than simply waiting for user commands; 2) the human-centered assumption: kids take initiative to discover and learn by themselves, while the computer agent plays auxiliary functions. The two assumptions require the computer to keep track of the users closely, detecting their status including the learning stages, cognitive activities, requests/inquiries, attention foci, and so on.

The characteristics of the project make mental activity detection an important cue for the computer to make response decisions. According to the characteristics of a classroom environment and the conditions under which the computer agent would involve response, the mental states of kids generally can be summarized into three cases: confidence, confusion and frustration. When the user is confident about his/her action, e.g., "I see the small one moving faster", the computer agent needs to initiate communication by providing suggestion for his/her next action. When the user is confused, e.g., "What do you mean?" the computer agent needs to help him/her get a better understanding or provide guidance by asking open questions. When the user is frustrated or hesitant, e.g., "I count eight then for the ... sort of one, but..." the computer agent needs to encourage or help the user clarify his/her ideas. According to the utterances collected in the experiments, the ratio of confidence: confusion: frustration = 0.6: 0.3: 0.1.

2. SYSTEM DESCRIPTION

2.1. System overview

The detection of mental states is based on the prosodic features and key words/phrases associated with semantic meaning. The pilot study reveals that children have somewhat different expressions than adults when interacting with computers. One difference is that children prefer short sentences to long sentences for expression, probably due to the small vocabulary size and shortage of linguistic knowledge. For example kids, especially boys, like to express their agreement and denial using a simple "hum" but with different tones. In this case, prosody provides the key cue to discriminate the user's intention.

However, some utterances are inherently ambiguous for mental state detection by means of prosody alone. For example, wh-questions, how-questions, and no-opinion statements (e.g., "I don't know") represent confusion, but may be uttered with prosody indistinguishable from the prosody of a confident utterance. Spotting some words/phrases embedded in the fluent speech helps discriminate the difference and make a correct decision. The integration scheme of prosodic and lexical information is depicted in Figure 1.

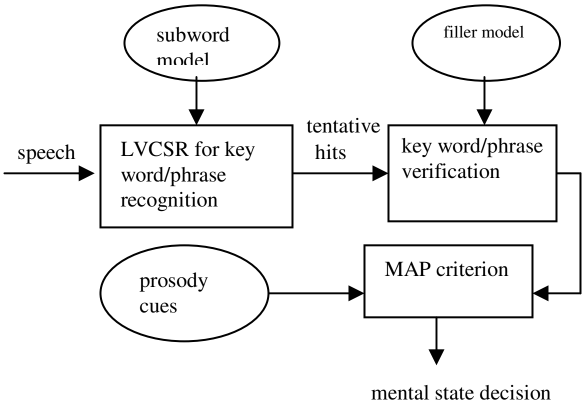

Figure 1. Overview of the mental state classifier.

2.2. Prosodic analysis

Summarized into Table 1, the prosodic features used in the mental state discrimination are divided into several types:

1) *Pitch*: F0 is derived using a typical autocorrelation method, but with the expected pitch range adjusted in order to account for the difference between children's speech and adult speech [1]. Confused children tend to ask more questions; furthermore, confused children tend to exaggerate the pitch-rise at the end of their questions, while confident children tend to exaggerate the turn-final declaration fall. Unlike either confidence or confusion, in some cases the pitch contour at the end of frustrated sentences is neither rising nor falling.

2) *Voiced-Unvoiced Percentage:* The time percentage that voiced speech occupies in an utterance is a good cue for discriminating frustration from confidence and confusion.
3) *Energy-related features:* Typically, an utterance falls to lower energy when close to completion. However, when speakers stop mid-stream, this fall has not yet occurred and thus energy remains unusually high [2]. Therefore, the energy can be an indicator of frustration.
4) *Pause-related features*: As a cue of frustration, pause refers to a time period of non-speech signals over 100ms. If the non-speech duration is less than 100ms, it is more likely to be a natural transition within sentences. To detect a pause, we use the time-varying threshold method of Li et al. [3].
5) *Syllabic-rate-related features*: Syllabic rate is defined as the number of syllables normalized by the utterance duration. Usually people speak more slowly when they are hesitant and confused, as opposed to when they are confident. The speaking rate is computed by the average of three estimators. The first estimator is peak counting performed on the wide-band energy envelope, and then normalized by the utterance duration. The second estimator is a sub-band-based module proposed by Morgan and Fossler-Lussier [4]. Our third estimator is the modulation-spectrum method of Kitazawa et al. [5].

Table 1: The prosodic features used in the mental state classification

Features	Description
F0_ratio	ratio of mean F0 over the end region (the final 100ms) and the penultimate region (the previous 100ms).
F0_reg_pen	least-square all-points regression over the penultimate region.
F0_reg_end	least-square all-points regression over the end region.
F0_norm	the number of frames with non-zero F0 in an utterance normalized by the utterance duration.
E_ratio	ratio of energy over the end region and the penultimate region
pau_norm	total pause durations normalized by the utterance duration.
syllarate	number of syllables normalized by the utterance duration.

2.3. Lexical information and word spotting

The detection of key words/phrases makes use of a LVCSR system. Sampled at 11 KHz, the speech input is pre-emphasized and grouped into frames of 330 samples with a window shift of 110 samples. The speech signal is characterized by 13 MFCC components normalized by cepstral mean subtraction, and log-scaled energy normalized by the peak. Moreover, their deltas and delta-deltas are also computed. Therefore, each speech frame is represented by a vector of 42 features.

The key word/phrase spotting comprises detection, and subsequent verification to reduce false alarms. The detection portion is based on LVCSR, in which each key word/phrase is represented by the concatenation of the models of its component phones. Each subword model is a left-to-right 3-state HMM with 16 Gaussian mixtures per state. The universal background subword models trained from TIMIT database are further adapted to children of various ages and sex. The speech adaptation is based on maximum likelihood linear regression (MLLR) followed by maximum a Posteriori (MAP) adaptation. Recognition is accomplished by a frame synchronous Viterbi search algorithm to determine the sequence of words that maximizes the likelihood of a given utterance.

The putative results are further validated by employing a filler model, based on the geometric log likelihood mean proposed by Sukkar and Lee [6]. The keyword acceptance/rejection is determined by comparing V(O; W) with a predefined threshold, where

$$V(O;W)=\sum_{j=1}^{N}\{\log[L(O_j \mid s_j)]- \log[\frac{1}{M_j}\sum_{m=1}^{M_j}\exp(\gamma\log[L(O_j \mid s_j(m))])]^{1/\gamma}\}, \quad (1)$$

and the key word W is the concatenation of N subwords;

$L(O_j|s_j)$ is HMM likelihood score for O_j, the corresponding observation sequence, given s_j, the j^th subword model;

M_j is the total number of subwords in the corresponding cohort set of s_j;

$s_j(m)$ is the m^th subword in the corresponding cohort set of s_j; and

γ is constant.

2.4. Uniqueness of children's speech

Children under 13 years old have very different acoustic characteristics depending on their age. The variability lies in two aspects: (1) age-dependent variability in terms of formants; (2) intra-speaker variability in terms of cepstral distance both within a token and across two repetitions [7]. 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. The variability results in degradation of ASR performance on children. It has been reported that the in-vocabulary word error rate for children is almost twice that of adult users. To compensate for these variabilities, we apply frequency warping to normalize vocal tract length [8]. The scheme of frequency warping is shown in Figure 2. The frequency *f* is warped by means of a bilinear rule, which maps an input to an equal length of output in the frequency domain.

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

where f_N is the Nyquist frequency, and β_f is the frequency-dependent warping factor.

To compensate for the inter-speaker variability, different warping factors are used on groups of children with the same age and sex. With reference to the data published in [1], for each group, the warping factors at formants F1, F2 and F3 are computed as the ratio of average formant values of that group to those of adult males.

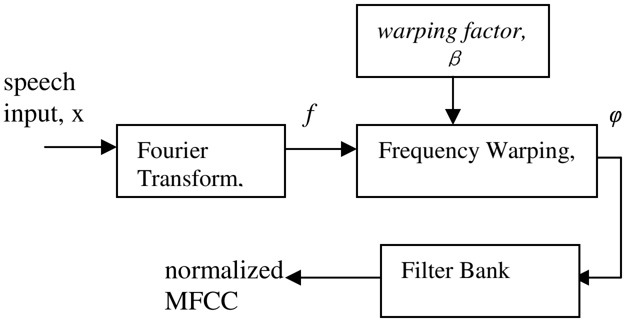

Figure 2. Vocal tract length normalization by frequency warping.

To compensate for the intra-speaker variability, we put forward that the warping factors for frequencies other than the three formants are approximated by interpolation. The interpolation scheme is shown in Figure 3, where the three critical points are the group-dependent (F_1, β_1), (F_2, β_2) and (F_3, β_3), respectively.

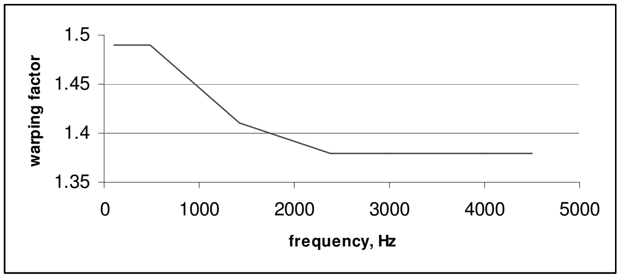

Figure 3. The interpolation pattern for deriving frequency-dependant warping factors to compensate intra-speaker variability.

2.5. Mental state classification

The lexical and prosodic information are integrated together for mental state detection. Denote the mental state as E, the recognized keywords or their combination set as W, and prosodic features as F. Then the a posteriori most probable hypothesized mental states, given the prosodic evidence and word identities, is derived by the MAP criterion

$$E^*=\arg\max_E P(E \mid W,F)=\arg\max_E P(W,F \mid E)P(E)$$

$$=\arg\max_E\{\log P(F \mid W,E)+\log P(W \mid E)+\log P(E)\}. \quad (3)$$

In the computation of *P(W|E)*, insufficient training samples mean that some possible cases might not be observed during training. To compensate for the inherent sparseness of data, we used a standard unit-discount smoothing technique [9]. That is, each of the training samples can be set to occur once more than it really does:

$$P(W_i \mid E) = \frac{1+T(W_i,E)}{\sum_{W_j} 1+T(W_j,E)}, \qquad (4)$$

where $T(W_i, E)$ is the times when W_i and E occur simultaneously.

3. EXPERIMENTAL RESULTS

In this domain-specific dialogue system, 21 key words/phrases were selected as being most relevant to the problem of mental state determination. The key words/phrases are listed in the following table.

Table 2: A list of key words/phrases used in the mental state classification

yeah/yes	no	because
I think	uhm	so
which	where	when
what/what's	I don't know	do you/I
can you/I/we	could you	this one
is it	why	these
I'm not sure	should I	how

Phoneme models were constructed using HTK, and trained using the TIMIT database. Models of children ages 7-9 were then adapted using the "CMU kids" database, a microphone speech data corpus distributed by the Linguistic Data Consortium, and provides substantial dictation speech of children within ages of 7 to 9. Phoneme models for children of other ages had to be adapted using a few minutes of utterances by the users themselves. The models of prosody feature distribution were trained using 300 sample utterances manually extracted from the data collected in the project experiments. The tests were performed on 50 test utterances collected in the same way. Of the 50 test utterances, 25 were spoken by an 11-year-old girl, and the other 25 were spoken by a 9-year-old boy. None of the training utterances included data from either of the two test users; models were adapted to the speech of the 11-year-old using her own speech, but models were adapted to the speech of the 9-year-old using speaker-independent data from the "CMU kids" corpus.

Three-class classification accuracies were 76% and 72% for the 11-year-old girl and the 9-year-old boy, respectively. Test results showed that background noise was an important reason for the classification error. In fact, noise affected the prosodic features derivation, and affected the key word/phrase recognition result. Therefore, the future work on noise reduction and speech enhancement is expected to improve the classification accuracy.

Sometimes the inexplicit utterance of children is another important source of classification error. Comparing the two kids in the experiment, the utterances of the 9-year-old boy are less explicit than the 11-year-old girl, resulting in more recognition errors in key word/phrase spotting.

4. CONCLUSION

Through the prosodic analysis and key word/phrase spotting of children's spoken language, the mental activities of children can be classified during their interaction with computers. The inter-speaker variability and intra-speaker variability of children's speech are compensated for by a vocal-tract-length-normalization (VTLN)-based technique. Tests on a set of 50 utterances collected from the project experiment showed the classification accuracy was 74%. The test results showed that noise is an important reason for the classification error. So the future work on noise reduction and speech enhancement is expected to improve the classification accuracy.

5. ACKNOWLEDGEMENTS

This work was supported by the NSF.

REFERENCES

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

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

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

[4] N. Morgan and E. Fosler-Lussier, "Combining multiple estimators of speaking rate," in *IEEE ICASSP*, 1998.

[5] S. Kitazawa, H. Ichikawa, S. Kobayashi, and Y. Nishinuma, "Extraction and representation rhythmic components of spontaneous speech," in *EUROSPEECH*, pp. 641-644, 1997.

[6] R. A. Sukkar and C.-H. Lee, "Vocabulary independent discriminative utterance verification for non-keyword rejection in subword based speech recognition," *IEEE Trans. Speech and Audio Proc.*, 4: 420-429, 1996.

[7] S. Narayanan and A. Potamianos, ``Creating conversational interfaces for children," *IEEE Trans. Speech and Audio Proc.*, 10(2): 65-78, 2002.

[8] P. Zhan, and A. Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition," *CMU Computer Science Technical Reports*, 1997.

[9] S. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, 13: 359-394, 1999.