Efficient Dialogue Management Of A Spoken Dialogue System By Using Prosody Information Of The User Utterances

Manzoor Ahmad*, Dr. S. M. K. Quadri**
*(P.G. Department of Computer Science, University of Kashmir, Srinagar J&K-190006)
** (P.G. Department of Computer Science, University of Kashmir, Srinagar J&K-190006

ABSTRACT

Words used by a human while framing a response during the interaction with a software agent like spoken dialogue system(SDS) has valuable information as regards internal mental state of the user is concerned. The users level of certainty about a response could be judged by the prosody information structure. Prosody reveals Information about the context by highlighting information structure and aspects of the speaker hearer relationship. Most often it is observed that the speakers internal state is not depicted by the words he uses but by the tone of his utterance or facial expression of the user. This paper uses PRATT a tool for speech analysis which uses 15 acoustic features to determine the certainty of the responses of the user using the prosody information which will actually aid the dialogue management component of the SDS in framing a better dialogue strategy.

Keywords - uncertainty handling , Prosody information , spoken language understanding , machine learning.

I. INTRODUCTION

Spoken language is an intuitive form of interaction between humans and computers. Spoken Language Understand has been a challenge in the design of the spoken dialogue system where the intention of the speaker has to be identified from the words used in his utterances. Typically a spoken dialogue system comprises a four main components an automatic speech recognition system (ASR) , Spoken language understanding component (SLU) , Dialogue manager (DM) and an Speech synthesis system which converts the text to speech (TTS). Spoken Language understanding deals with understanding the intent from the words of the speakers utterances. The accuracy of the speech recognition system is questionable and researchers have provided various solutions to the problem of automatic speech recognition which lagged behind human performance [2],[3] there have been some notable recent advances in discriminative training [4]; e.g., maximum mutual information (MMI) estimation [5], minimum classification error (MCE) training [6], [7], and minimum phone error (MPE) training [8], [9]), in large-margin techniques (such as large margin estimation [10], [11], large margin hidden Markov model (HMM) [12], large-margin MCE [13]–[15], and boosted MMI [16]), as well as in novel acoustic models (such as conditional random fields (CRFs) [17]–[19], hidden CRFs [20] ,[21] and segmental CRFs [22]) ,training densely connected, directed belief nets with many hidden layers which learn a hierarchy of nonlinear feature detectors that can capture complex statistical patterns in data [23]. There are many cases of experiences by the users when the computers either do not understand the intended meaning of the user even after correctly recognizing the spoken utterances. One of the reason may be that in a face to face human conversation , there are contextual, audio and visual cues [1] which aid the knowledge requirements of the users for the efficient communication as the users other than the contextual are able to sense the mood and tone of the user by which they come to know whether the speaker is certain or not. This is , absent in a dialogue between a computer and a human because in many potential applications there is only audio input and no video input. If the Spoken Dialogue Systems are improved to use the prosodic information from the spoken utterance they will definitely benefit from the level of certainty of the user [24] such as spoken tutorial dialogue systems[25] , language learning systems [26] and voice search applications [27] . Our primary goal is to make use of prosodic information for aiding the dialogue manager in selecting the dialogue strategy for effective interaction and influencing the final outcome. Technically Prosody is defined as the rhythm, stress, and intonation of speech which reflect various features such as emotional state of the speaker , the form of the utterance (statement, question, or command , the presence of irony or sarcasm, ; emphasis, contrast, and focus or other elements of language that may not be encoded by grammar or choice of vocabulary Prosodic information of an utterance can be used to determine how certain a speaker is and hence the internal state of mind [28] which can be used for tasks from detecting frustration[29] , to detecting flirtation [30] and other intentions. The model proposed that uses prosodic information to classify utterances has effectively colored the system responses in a travel based information system and performed better than a trivial non-prosodic baseline model.

In the context of human computer interaction , the study of prosodic information has been aimed at extracting mood features in order to be able to dynamically adapt a dialog strategy by the automatic SDS.

II. Corpus and Certainty annotation.

It is very important to understand that not only what words are spoken by a speaker in his utterance but how the words are spoken along with the certainty factor can actually guide the dialogue process between the machine and the user. The spoken utterance may be perceived as uncertain , certain , neutral or mixed which helps the dialogue system to make a guess about the mental state of the user about the utterance or about the concept about which he is speaking about. In this paper we examine manifestations of a travel desk attendant and a tourist certainness as it is expressed within the context of a spoken dialogue.

AGENT : How many days do you want me to plan your tour.
Tourist : Four to Five (UNCERTAIN)
AGENT : Is it Four or Five days.
TOURIST : Five days (CERTAIN)
AGENT : Would you like to visit snowy destination.
TOURIST : Uh-uhh (NEUTRAL)

Fig 1 . An annotated excerpt from the travel corpus.

A corpus of 100 travel related dialogs are selected and after listening each sentence of the tourist is labeled by an annotator with either certain or uncertain or neutral. The dialog were also lexically annotated based on the words used as certain , uncertain and neutral. The percentage of sentences with certainty , uncertainty and neutral for the auditory and lexical conditions are shown in the table 1.

Condition	Certain	Uncertain	Neutral
Auditory	22.3%	18.4%	59.3%
Lexical	12.1%	11.7%	76.2

Table 1 : Percentage of corpus with different levels of certainty , annotated by listening to the audio of the dialog context and annotated based on the lexical structure of the dialogues.

It was observed that 40.7% non-neutral corpus could be decided as certain or uncertain based on the audio and the dialog context compared to the 23.8% based on the lexical information. As such we used the acoustic-prosody features for further information about the certainty or uncertainty.

III. Prosody Model

For the basic model we compute values for 15 prosodic features as given in the table 1 for each utterance in the corpus of the travel data set using PRATT (a program for speech analysis and synthesis) [34] and WAVESURFER for extracting the f0 contour. . Feature values are represented as zscores normalized by speaker . The temporal features like speaking rate , Total silence, Percent silence, Speaking duration , Total duration are not normalized.

No. of features	Features
6	Mean absolute slope(Hz) , minimum , maximum and standard deviation, relative position min f0 , relative position max f0 statistics of fundamental frequency (f0) Pitch
4	Minimum , Maximum , Mean and Standard Deviation (RMS) , statistics of Intensity
1	Ratio of voiced frames to total frames in the speech signal as an approximation of speaking rate
2	Total silence, Percent silence.
2	Speaking duration , Total duration.

Table 2 : Extracted and selected features.

The set of features were selected in order to be comparable with Liscombe et al [31] who used the same features along with turn related features for classifying uncertainty.

IV. Classification Results

The features extracted are used as input variables to WEKA machine learning software which built C4.5 decision tree models boosted using AdaBoost which iteratively builds weak models and combines them to form a better model to predict the classification of unseen data. As an initial model we train a single decision tree using the selected 15 features as listed in table 2. The model was evaluated over all the utterances of the corpus and it classified within the classification classes , certain , uncertain and neutral with an accuracy of 64.23% as compared to the non-prosodic model which had a an classification accuracy of 51.1%

V. Conclusion

In human computer interaction , the computer have to act human like so that other than the lexical information, computer should be able to utilize the auditory and visual cues so that the users are responded in a manner which is based on their emotions and the system looks more user friendly. In an automated travel desk when a tourist has limited number of days and more destinations . selecting few based on his information and preference can help the automated SDS to design a travel plan which is more based on the preferences and prosodic information of the user . When the system talks about snow the prosodic features can indicate how much certain the traveler is about visiting destinations which contain more snow. Thus prosodic information provides information regarding the internal state of mind of the user and would help the dialogue manager to dynamically select the strategy based on the certainty or uncertainty.

In our experiment we used a small set of prosodic features that have been examined in related work by other researchers . Using and expanded set of features would improve the results and the accuracy with which the certainty can be detected. In the future work we would be using the visual cues like facial expressions , body language and other inputs by a human to maximize the ability to determine the internal mental state of the user which can give the spoken dialogue system a mechanism to select dynamic dialogue strategy.

REFERENCES

[1] E. Krahmer and M. Swerts, "How children and adults produce and perceive uncertainty in audiovisual speech," *Language and Speech*, 48(1), 2005, 29–53.

[2] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C. Lee, N. Morgan, and D. O'Shaugnessy, "Research developments and directions in speech recognition and understanding, part 1," *IEEE Signal Processing Magazine, vol. 26, no. 3*, pp. 75–80, 2009.

[3] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C. Lee, N. Morgan, and D. O'Shaugnessy, "Research developments and directions in speech recognition and understanding, part 2*," IEEE Signal Processing Magazine, vol. 26, no. 4*, pp. 78–85, 2009.

[4] X. He, L. Deng, and W. Chou, "Discriminative learning in sequential pattern recognition — a unifying review for optimization-oriented speech recognition," *IEEE Signal Processing Magazine, vol. 25, no. 5, pp*. 14– 36, 2008.

[5] S. Kapadia, V. Valtchev, and S. J. Young, "MMI training for continuous phoneme recognition on the TIMIT database," *In Proc. ICASSP, vol. 2*, 1993, pp. 491–494.

[6] B. H. Juang, W. Chou, and C. H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing, vol. 5, no. 3*, pp. 257–265, 1997.

[7] E. McDermott, T. Hazen, J. L. Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large vocabulary speech recognition using minimum classification error," *IEEE Transactions on Speech and Audio Processing, vol. 15, no. 1*, pp. 203–223, 2007.

[8] D. Povey and P. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," *in Proc. ICASSP*, 2002, pp. 105–108.

[9] D. Povey, "*Discriminative training for large vocabulary speech recognition,"* Ph.D. dissertation, Cambridge University Engineering Dept, 2003.

[10] X. Li, H. Jiang, and C. Liu, "Large margin HMMs for speech recognition," *in Proc. ICASSP*, 2005, pp. 513–516.

[11] H. Jiang and X. Li, "Incorporating training errors for large margin HMMs under semi-definite programming framework," *in Proc. ICASSP, vol. 4,* 2007, pp. 629–632.

[12] F. Sha and L. Saul, "Large margin gaussian mixture modeling for phonetic classification and recognition*," in Proc. ICASSP*, 2006, pp. 265–268.

[13] D. Yu, L. Deng, X. He, and A. Acero, "Use of incrementally regulated discriminative margins in MCE training for speech recognition," *in Proc. ICSLP*, 2006, pp. 2418–2421.

[14] D. Yu and L. Deng, "Large-margin discriminative training of hidden Markov models for speech recognition," in Proc. ICSC, 2007, pp. 429– 436.

[15] D. Yu, L. Deng, X. He, and A. Acero, "Large-margin minimum classification error training for large-scale speech recognition tasks," in Proc. ICASSP, vol. 4, 2007, pp. 1137–1140.

[16] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature space discriminative training," *in Proc. ICASSP*, 2008, pp. 4057–4060.

[17] Y. Hifny and S. Renals, "Speech recognition using augmented conditional random fields," *IEEE Transactions on Audio, Speech & Language Processing, vol. 17, no. 2*, pp. 354–365, 2009.

[18] J. Morris and E. Fosler-Lussier, "Combining phonetic attributes using conditional random fields," in Proc. Interspeech, 2006, pp. 597–600.

[19] G. Heigold, "A log-linear discriminative modeling framework for speech recognition," PhD Thesis, Aachen, Germany, 2010.

[20] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random

fields for phone classification," in Proc. Interspeech, 2005, pp. 1117–1120.

[21] D. Yu and L. Deng, "Deep-structured hidden conditional random fields for phonetic recognition," in Proc. Interspeech, 2010, pp. 2986–2989.

[22] G. Zweig and P. Nguyen, "A segmental conditional random field toolkit for speech recognition," in Proc. Interspeech, 2010.

[23] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation, vol. 18*, pp. 1527–1554, 2006.

[24] Heather Pon-Barry, Stuart M. Shieber , "Recognizing Uncertainty in Speech", *EURASIP Journal on Advances in Signal Processing, Volume 2011*

[25] K. Forbes-Riley and D. Litman, "Adapting to student uncertainty improves tutoring dialogues," Frontiers in Artificial Intelligence and Applications, vol. 200, no. 1, pp. 33–40, 2009.

[26] A. Alwan, Y. Bai, M. Black et al., "A system for technology based assessment of language and literacy in young children: the role of multiple information sources," in Proceedings of the 9th IEEE International Workshop on Multimedia Signal Processing (MMSP '07), pp. 26–30, Chania, Greece, October 2007.

[27] T. Paek and Y.-C. Ju, "Accommodating explicit user expressions of uncertainty in voice search or something like that," in Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH '08), pp. 1165–1168, Brisbane, Australia, September 2008.

[28] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," IEEE Transactions on Speech and Audio Processing, vol. 13, no. 2, pp. 293–303, 2005.

[29] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human computerdialog," in Proceedings of the International Conference on Spoken Language Processing, pp. 2037–2040, Denver, Colo, USA, 2002.

[30] R. Ranganath, D. Jurafsky, and D.McFarland, " It's not you it's me: detecting flirting and its misperception in speed-dates," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '09), pp. 334–342, Singapore, 2009.

[31] J. Liscombe, J. Hirschberg, and J. J. Venditti, "Detecting certainness in spoken tutorial dialogues," in Proceedings of the 9th European Conference on Speech Communication and Technology, pp. 1837–1840, Lisbon,Portugal, September 2005.

[32] K. Forbes-Riley, D. Litman, and M. Rotaru, "Responding to student uncertainty during computer tutoring: an experimental evaluation," in Proceedings of the 9th International Conference on Intelligent Tutoring Systems (ITS '08), pp. 60–69, Montreal, Canada, June 2008.

[33] J. C. Acosta and N. G. Ward, "Responding to user emotional state by adding emotional coloring to utterances," in Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH '09), pp. 1587– 1590, Brighton, UK, September 2009.

[34] P. Boersma , "Pratt , a system for doing phonetics by computer ". Glot International, vol 5 no. 9/10, pp.341-345, 2001