http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet

Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 7 (2): 366-369, 2008
ISSN 1812-5638

Enhanced Intentness Estimation in a Colloquy

[1]M. Nachamai, [2]T. Santhanam and [3]M. Muthuraman
[1]Mother Teresa Women's University, Kodaikannal, India
[2]Department of Computer Applications PG and Research,
DG Vaishnav College, Arumbakkam, Chennai, India
[3]Institute for Circuit and System Theory, University of Kiel, Germany

Abstract: This study proposes a methodology to find the interest levels of two speakers in a conversation. The ANN-HMM approach-a hybrid method is adopted. The hybrid method uses language input as an additional parameter in addition to the acoustic features. The language input provides a measure of classification of the input speech utterance. A combined classifier is used to make a linear decision on the emotion of the uttered speech as an arousal or valence. When the decision is fed to the Generative Factor Analyzed Hidden Markov Model (GFA-HMM) it evidently substantiates to be a better method with good accuracy rate of classification of whether the speaker is entangled in the conversation or vice-versa. The proposed method produced highly satisfactory results for the Linguistic Data Consortium (LDC) emotional prosody dataset.

Key words: Artificial Neural Network (ANN), generative factor analyzed hidden Markov model (GFA-HMM), minimum distance classifier (MDC), principal component analysis (PCA), linear discriminant analysis (LDA), MPL

INTRODUCTION

Automatic speech recognition systems generally use either statistical pattern recognition systems or knowledge based principles in their algorithms. In this paper the method deployed is a combination of both statistical pattern recognition as well a knowledge based approach (phonetic and linguistic). Previous studies advocate that different acoustic modeling technique will provide a better result with HMM to recognize emotions (Nogueiras *et al*., 2001).The recently proposed strategy overcomes the basic weakness of the considerable amount of time taken by HMM for recognition. The method decouples the main search with parallelization, so all the probability alignments are calculated more efficiently and quickly (Steward, 2004). Emotionally salient keywords may yield a better result for identifying emotion category (Lee and Narayanan, 2005). Grouping of independent language information from different documents, leads to be a good attribute to denote the overall findings of the document set (Haraty and Ariss, 2007). An analytical method of phonetic labeling is used instead of the global method. The study focuses on finding the interest or involvement levels of two speakers in a conversation-as non-negative if the speaker is interested in the conversation or as negative if the speaker is not interested. The methodology followed is novel in the following ways: First, the speech signal is taken as a band filtered input considering both the power spectrum and phase spectrum of each segment which accounts for the robustness of the method, Second, speaker recognition is normally done using only acoustic feature here we have considered an additional input of language information which gives the base salience measure of an emotion state (negative/non-negative) (Yu *et al*., 2004), Third, decision level fusion is adopted thus decreasing the complexity of the feature vector dimension by using a combined classifier, Fourth, GFA-HMM is used to get the encoded state sequences which proves to be a good performer than the existing isolated mathematical models. The general architecture is depicted in Fig. 1.

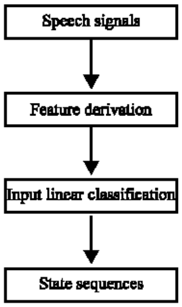

Fig. 1: General architecture

Corresponding Author: M. Nachamai, Mother Teresa Women's University, Kodaikannal, India

MATERIALS AND METHODS

The methodology implemented in the study is an amalgamation of statistical approach with an analytical method. The algorithm is split into three phases, First phase, tells about the Feature Derivation process from the input data, Second phase, explicates the Linear Categorization through a combined classifier, Third phase; depict implementation of GFA-HMM producing State Sequence Generation identifying intent levels.

Feature derivation: Feature extraction plays a major role in the affective computation process. There are enormous speech features that can be extracted from the speech signal but the problem lies in identifying which is the best feature that affects the results and the methods to be adopted to make those inputs robust for any situation. When we consider the problem of conversation between two speakers it is going to be exchange of dialogs which is encompassed of more pauses, overlaps and gaps than in the normal speech recognition data. The work here considers the power spectrum and phase spectrum for each band filtered input speech segment. When the power spectrum alone is considered it is not robust to the environmental noise and distortion (Paliwal and Atal, 2003). The phase spectrum is vulnerable for any environmental disturbed input data. From these segments two inputs are to be derived:

- Acoustic information
- Language information

Linguistic units considered for acoustic information are sub word units. The sub word units are more accurate than the other acoustic features (Wildermoth and Paliwal, 2004). The phonemes and di-phones are related to each sub word unit through a lexicon and a word network. A word network is included along with the lexicon for spotting the out-of-vocabulary words in the conversation. The lexicon used in the paper is WordNet where the words are organized on a lexical concept. The use of lexical resources gives a better performance (Plas *et al*., 2004). The lexicon helps in auto assignment of keywords to multi party dialogue. The scoring method applied in the paper for the lexicon is based on the connectivity and concept reiteration. Connectivity is measure between the words in the concept they appear. The general method used is the word reiteration, here we have taken concept reiteration since the work requires the analysis of the concept in which the speakers converse.

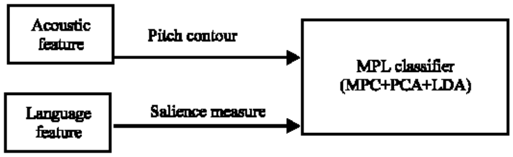

Fig. 2: Feature extraction

The feature set includes acoustic-prosodic as well as linguistic information. From the speech segment base features are to be derived, from the base features pitch signal for each lexical unit is alone considered which attributes to the entire feature space. From each sub-word unit the pitch contour is measured (Fig 2). Depicts the feature extraction process. The multiple input in the method accounts for incomplete and missing data. The reason for assimilating two inputs is to avoid the discrepancy in the final accuracy. In a colloquy, speaker 1 may remain silent on listening to the speaker 2 where the state of speaker 1 does not change, where no state change happens it cannot be labeled as a negative emotion, when both the pitch contour data and the emotion measure is calculated together in such a situation will account for a good decision. The pitch contour data for a segment remains same when a speaker stays listening. So if the state change is not shown the previous state probability is considered which gives a good credential on the emotion state decision of the speaker.

Combined classifier-MPL: The two inputs when combined and directly fed to the network will have a dimension crisis and the feature vector derived as a combination may hinder the performance or efficacy of the neural network. A better strategy is followed by taking a decision fusion at the input level. Both the inputs are fed to the classifier which produces the combined feature vector thus reducing the constraint on dimension. The combined classifier used here is a linear classifier (Sharma *et al*., 2006). The MPL classifier used in the work is a linear combination of Minimum Distance Classifier (MDC), Class- Dependent Principal Component Analysis (PCA) and a Linear Discriminant Analysis (LDA). This classifier is capable of making an accurate linear classification give an emotive class output. This classifier supersedes the normal isolated linear classifiers on both aspects of accuracy and storage.

State sequence generation using GFA-HMM: The input feature vector that is the emotive class output from the MPL is fed to the GFA-HMM. The output observation vector is correlated with the pitch feature which is considered as the acoustic feature unit which is a dependent continuous latent vector. The correlation is done at every particular time interval -t, where the density of the pitch feature is dependent on the state of the HMM at time t. These correlations are then derived into an independent measure of acoustic unit, which is a distributed standard Gaussian variate. Conventional methods of HMM use only a diagonal covariance matrix for each inter-frame, represented by a Mixture of Gaussians (MoG) which is discrete value dependent. GFA-HMM is considered since it works well with a limited number of parameters (Yao *et al*., 2003). In this work only a single feature of the acoustic unit is considered. GFA considers a latent form of representation at level one and in the next level it considers acoustic unit independent representation which proves to be a major advantage. It has a compact representation of intra-frame statistics which proves good for two reasons one is minimized dimension thus reducing the complexity of calculations and accounting for state change information accurately. An Expectation-Maximization algorithm is written for GFA-HMM with the maximum likelihood at each state. Since the work is on a conversation involving spontaneous speech, the method adopted must be able to balance pauses in between speech, hesitations, false starts, overlapping of speech signals and ungrammatical constructions.

The GFA-HMM calculates four different measures:

- Participation probability - probability of state change of a speaker (negative to non-negative/ non-negative to negative)
- Speaker probability - probability of speaker being in a state at time t

These two probabilities are calculated for both speaker 1 and speaker 2 and given as an encoded state sequence. The state sequences are decoded with the viterbi search algorithm assigning the speaker involvement on a 1-5 scale grading.

$$D(R_i, b, e) = D'\,(R_{iw}, e) - D'\,(R_{i\varepsilon}\text{-}b) \tag{1}$$

As given in the Eq. 1, instead of taking only the local distortion measure the viterbi search algorithm takes the cumulative distortion for the word R beginning at b and ending at e is calculated by taking into account the local distortion for the observation sequence in state i. This yields a better search vector when applied to the phoneme like template model for automatic speech recognition (Riley and Litman, 2004). The viterbi algorithm searches for the best word in sequence with the phonetic probabilities calculated with the time change. The decoded sequence turns to the emotion evolved by the speaker in the conversation.

Table 1: Accuracy rate of colloquy classification

Random (%)	Isolated SVM (%)	HMM (%)	Coupled HMM (%)	GFA-HMM (%)
20	47	61	63	78

RESULTS AND DISCUSSION

Real world data corpuses are not available with large emotion types or vocabulary information. The chosen data set LDC Emotional Prosody corpus consists of 14 types of emotion. Each emotion type has 25 spoken utterances by actors/ actresses of which 13 utterances were used for training and the rest for testing purposes. The outcomes of detecting speakers involvement in a conversation is shown in Table 1. The proposed method has yielded satisfactory result when compared with methods reported in (Wiebe *et al*., 2005). The upshot is an efficient proof for the combinatorial approach of a statistical pattern recognition method with a knowledge based principle of linguistic information and a dynamic model of an ANN. A mixture of a Gaussian method proves good for the non-linear features of speech signals.

The GFA-HMM has produced an accuracy rate of 78% which is realistic as opposed to the traditional methods for classification in speech recognition problems.

CONCLUSION

The emotion detection problem dealt in this article is a pioneering approach that paves the way for many new solutions in affective computing. The technical contribution of the work is emotive detection, working in a conversation and not based as usual identifying emotions of an individual speaker. The results clearly demonstrate the usage of multiple inputs, robustness considered and the methodology employed. The experimental data employed is an enacted corpus and hence may not prove to be the same for real time speech problems. The non availability of real colloquy with large vocabulary is a major problem. Use of more training data with many discrete emotions on real speech data is a direction for future work.

REFERENCES

Haraty, R.A. and O.A. Ariss, 2007. CASRA+: A colloquial arabic speech recognition application. Am. J. Applied Sci., 4 (1): 23-32.

Lee, C.M. and S.S. Narayanan, 2005. Towards detecting emotions in spoken dialogs. IEEE. Trans. Speech Audio Process., 13: 2.

Nogueiras, A., A. Moreno, A. Bonafonte and J. Mariñño, 2001. Speech emotion recognition using Hidden Markov Models. In: Proc. Eurospeech. Aalborg, Denmark.

Paliwal, K.K. and B.S. Atal, 2003. Frequency- related representation of speech. In: Proc. EUROSPEECH., Geneva.

Plas, L.V.D., V. Pallotta, M. Rajman and H. Ghorbel, 2004. Automatic keyword extraction from spoken text-a comparison of two lexical resources: The EDR and wordnet. In: Proceedings of the LREC 2004 International Conference, Lisbon, Portugal, pp: 26-28.

Riley, K.F. and D.J. Litman, 2004. Predicting emotion in spoken dialogues using multiple sources. In: Proc. ACL.

Sharma, A., K.K. Paliwal and G.C. Onwubolu, 2006. Class-Dependent PCA, MDC and LDA: A combined classifier for pattern classification. Pattern Recog., 39 (7): 1215-1229.

Steward, A., 2004. A fast HMM match algorithm for very large vocabulary speech recognition. Speech Commun., 42 (2): 191-206.

Wiebe, J., T. Wilson and C. Cardie, 2005. Annotating expressions of opinions and emotions in language. LRE., 39 (2-3): 165-210.

Wildermoth, B.R. and K.K. Paliwal, 2004. Speaker recognition using acoustically derived units. In: Proceedings Microelectronic Engineering Research Conference, Griffith University.

Yao, K., K.K. Paliwal and T.W. Lee, 2003. Speech recognition with a generative factor analyzed hidden markov model. In: Proceedings EUROSPEECH, Geneva.

Yu, C., P.M. Aoki and A. Woodruff, 2004. Detecting user engagement in everyday conversations. In: Proceedings of the 8th International Conference on Spoken Language Processing, Jeju, Korea.