World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:12, 2007

Automatic Recognition of Emotionally Coloured Speech

Theologos Athanaselis, Stelios Bakamidis, and Ioannis Dologlou

***Abstract*—**Emotion in speech is an issue that has been attracting the interest of the speech community for many years, both in the context of speech synthesis as well as in automatic speech recognition (ASR). In spite of the remarkable recent progress in Large Vocabulary Recognition (LVR), it is still far behind the ultimate goal of recognising free conversational speech uttered by any speaker in any environment. Current experimental tests prove that using state of the art large vocabulary recognition systems the error rate increases substantially when applied to spontaneous/emotional speech. This paper shows that recognition rate for emotionally coloured speech can be improved by using a language model based on increased representation of emotional utterances.

***Keywords*—**Statistical language model, N-grams, emotionally coloured speech

I. Introduction

RECOGNISING the verbal content of emotional speech is a difficult problem, and recognition rates reported in the literature are in fact low. Although knowledge in the area has been developing rapidly, it is still limited in fundamental ways. The first issue concerns that not much of the spectrum of emotionally coloured expressions has been studied. The second issue is that most research on speech and emotion has focused on recognising the emotion being expressed and not on the classic Automatic Speech Recognition (ASR) problem of recovering the verbal content of the speech. Read speech and non-read speech in a 'careful' style can be recognized with accuracy higher than 95% using the state-of-the-art speech recognition technology. Including information about prosody improves recognition rate for emotions simulated by actors, but its relevance to the freer patterns of spontaneous speech is unproven.

Phonetic descriptions of emotional speech show that it has multiple features which would be expected to pose problem for ASR systems. Five areas of difficulty stand out. 1) Source [1], 2) Intensity [2], 3) Speech quality [3], 4) Prosody [4], 5) Timing [5].

A solution to the problem of emotional speech recognition is to modify the training process so that recognition is sensitive to prosodic information. Polzin & Waibel [6] show that this strategy can be effective. This paper deals with a second strategy, which is complementary to Polzin & Waibel's. It is well known that the emotion affects language as well as speech variables. For that reason the important issue is to identify corpora that reflect emotion-influenced language so that emotion-oriented language models can be learned from them. The language models are derived by adapting an already existing corpus, the British National Corpus (BNC). An emotional lexicon is used to identify emotionally coloured words, and sentences containing these words are recombined with the BNC to form a corpus with a raised proportion of emotional material.

This paper confirms that emotion does have major effects on recognition rate. The aim of this paper is to investigate the performance of a speech recognition system which is based on emotional oriented language model, for material that presents emotion variability. For experimental purposes a set of 4 different emotional characters are used.

The paper is organized as follows: the architecture of the speech recognition engine in section 2. The 3rd section describes the basic language model while in the next section follows a detailed presentation of the enhanced language model generation. The experimental scheme and results of using the basic and the enhanced language model are discussed in section 5 and concluding remarks are made in section 6.

II. System Architecture

The proposed large vocabulary continuous speech recognition system is based on Hidden Markov Models (HMM) [7]. The unknown speech input is converted into a sequence of acoustic vectors $Y = y_1, y_2, ..., y_n$, by means of a parameter extraction module. The goal of the LVR system is to determine the most probable word sequence $\hat{W}$ given the observed acoustic signal Y , based on the Bayes' rule for decomposition of the required probability $P(W \mid Y)$ into two components, that is,

Manuscript received January 8, 2006. Theologos Athanaselis is with the Institute for Language and Speech Processing, Artemidos 6 and Epidavrou, Maroussi, Athens, Greece, GR-15125, (phone: +302106875416; fax:+302106854270; e-mail: tathana@ilsp.gr).

Stelios Bakamidis, is with the Institute for Language and Speech Processing, Artemidos 6 and Epidavrou, Maroussi, Athens, Greece, GR-15125, (e-mail: bakam@ilsp.gr).

Ioannis Dologlou is with the Institute for Language and Speech Processing, Artemidos 6 and Epidavrou, Maroussi, Athens, Greece, GR-15125, (e-mail: ydol@ilsp.gr).

$$\hat{W} = \arg\max_{W} P(W/Y) = \arg\max_{W} \frac{P(W)P(Y/W)}{P(Y)} \quad (1)$$

The prior probability $P(W)$ is determined directly from the language model. The likelihood of the acoustic data $P(Y|W)$ is computed using a composite HMM representing W constructed from simple HMM phoneme models joined in sequence according to word pronunciations stored in a dictionary. Figure 1 illustrates the main components of the speech recognition module.

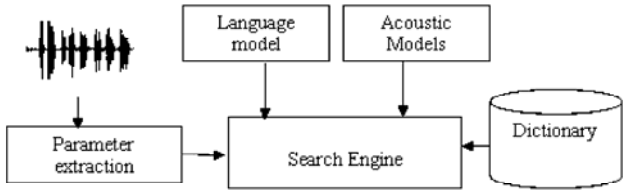

Fig. 1 The architecture of the recognition system

III. LANGUAGE MODEL

The language model (LM) that is used subsequently is the standard statistical N-grams [7]. The N-grams provide an estimate of $P(W)$, the probability of observed word sequence W. Assuming that the probability of a given word in an utterance depends on the finite number of preceding words, the probability of N-word string can be written as:

$$P(W) = \prod_{i=1}^{N} P(w_i \mid w_{i-1}, w_{i-2}, \ldots, w_{i-(n-1)}) \quad (2)$$

IV. ENHANCED LANGUAGE MODEL

When recognizing emotional speech, it is necessary to deal with linguistic phenomena that are not encountered in read speech. Although these do not affect human speech understanding, they lower the performance of speech recognition systems. This paper incorporates an algorithm [8] to improve the recognition rate by using an emotionally enhanced language model. To do so emotional text is extracted from the BNC using the Whissell emotional dictionary [9]. The Whissell dictionary comprises approximately 8700 words with emotional meaning. Here a subset of 2000 words of the Whissell lexicon is used. These words are the most frequent words of BNC that also belong to the Whissell list. The emotionality of a speaker's utterance affects both the prosodic parameters and the content. As a convenient way to model the effect on content, the existing BNC is enhanced by including emotional sentences. The enriched corpus is then used for language model design.

The first step is to extract the sentences from BNC that their component words belong to sub-Whissell dictionary. The Whissell corpus consists of these sentences. Next, the Whissell corpus is appended to the BNC λ times in order to create an emotionally enriched text corpus (emotional corpus). This corpus is used to train the emotionally enhanced language model. The factor λ is adjusted experimentally to maximise recognition performance.

The following formula depicts the merge of two different corpora in order to generate an emotional corpus:

$$S_{E.C} = S_{BNC} + \lambda \cdot S_{Whissel} \quad (3)$$

where, S_{BNC} is the number of sentences of BNC, $S_{Whissel}$ refers to the number of sentences of Whissell corpus, λ is the factor, and the total number of sentences is described by $S_{E.C}$.

From previous work [8] has been experimentally established that best results are obtained for λ=10. The BNC contains about 6.25M sentences. The Whissell's corpus has 0.3M sentences. The emotional corpus has $9^{1/4}$M sentences; this figure is derived by Equation 3 [8].

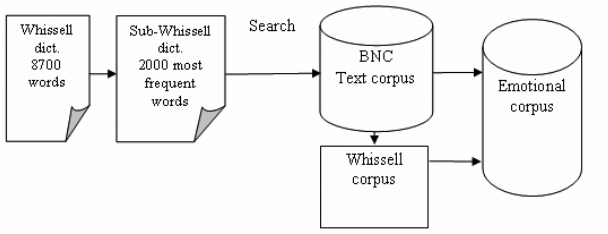

Fig. 2 The schematic view of text corpus enrichment using emotional sentences extracted from BNC corpus according to the sub-Whissell dictionary

V. PERFORMANCE OF THE SYSTEM

The experimentation involves speech recogniser with the basic language model and a test-set of sound files with emotional utterances.

The sound files were drawn from a database of spontaneous emotionally coloured speech developed as part of the ERMIS project (IST-2000-29319) [10]. The speech is produced by people holding conversations with SAL [11], a system which responds to speakers with emotionally coloured stock phrases. Speakers have to engage voluntarily with the exercise, but once they do, the system's contributions allow an emotionally charged atmosphere to be maintained, and encourage speakers to explore a range of emotional tones. The result is by some way the largest available database of spontaneous emotionally coloured speech, totalling over 5 hours. About half of it, involving four speakers, is recorded with a sound quality that allows ASR. Speakers' emotional state is assessed by trained raters using the Feeltrace system[11],[12], which raters use to indicate where they consider a speaker lies in a space with two

dimensions, activation (from highly energetic to torpid) and evaluation (from very positive to very negative). Ratings are made by using an on-screen cursor to 'track' the speaker's perceived emotional state in real time, producing a record of perceived emotional state in the form of two continuous traces, one for activation level, the other for evaluation. The database includes traces from four raters. This study used the one whose ratings were judged most reliable [13].

In order to evaluate speech recognition performance according to different emotional states ("PRAGMATICS", and "ANGRY") 100 different sound files for each speaker were used (50 sound files per emotional state).

A. Results with basic model

Figure 3 presents the percentage of correctly recognised words (y-axis) against the emotional states (x-axis), using different speakers. The corresponding values for different speakers are depicted by the legend with variations in grey-scale color. In the case of emotional state "Pragmatics", the lowest percentage of correctly recognised words is 13.4%, for speaker E, while the highest percentage of correctly recognised words is 33%, for speaker R. On the other hand, in the case of emotional state "Angry", the lowest percentage of correctly recognised words is 14%, for speaker E, while the highest percentage of correctly recognised words is 50%, for speaker R.

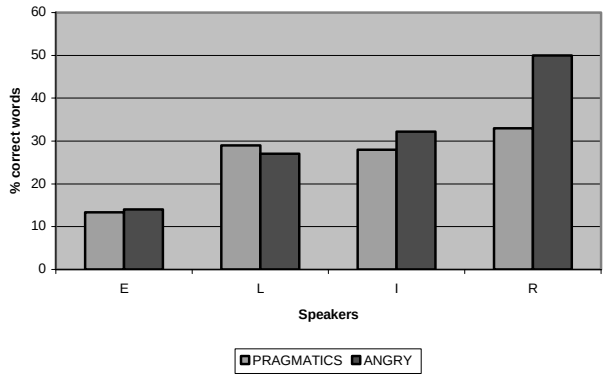

Fig. 3 The percentages of correctly recognised words for different emotional states and for different speakers, using a basic language model

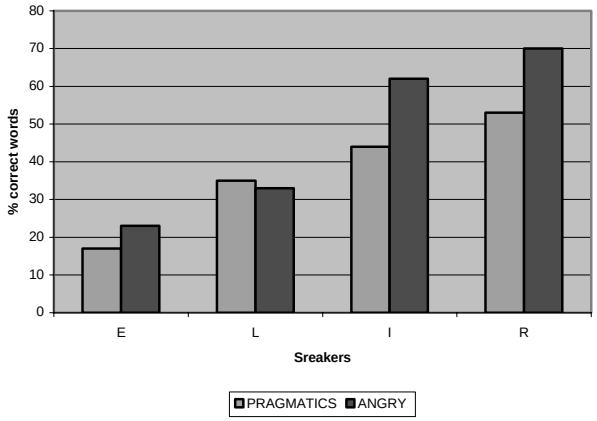

Fig. 4 The percentages of correctly recognised words for different emotional states and for different speakers, using an enhanced language model

B. Results with the enhanced model

Figure 4 presents the percentage of correctly recognised words (y-axis) against the emotional states (x-axis), using the enhanced language model. It is obvious that the percentages of correctly recognized words increase for all speakers and in each emotional state. This can be explained assuming that the speech recognition machine is trained by emotional data rather than general interest data.

VI. Conclusions

Recognising the verbal content of spontaneous emotionally coloured speech is a very difficult task. The results indicate that the speech recogniser has better performance enhancing the emotional characteristics of the language model. In general, it is noted that the percentage of correctly recognised words rises independently of the emotional state.

References

[1] K. Cummings, and M. Clements, Analysis of the glottal excitation of emotionally styled and stressed speech. *JASA,* 98 (1), pp 88-98, 1995.

[2] H.J.M. Steeneken, and J.H.L. Hansen, Speech Under Stress Conditions:Overview of the Effect of Speech Production and on System Performance. *IEEE ICASSP-99: Inter. Conf. on Acoustics, Speech, and Signal Processing* 4, pp 2079-2082, 1999.

[3] R. Cowie, and R. Cornelius, Describing the Emotional States that are Expressed in Speech. *Speech Communication,* 40, pp 5-32, 2003.

[4] D.J. Litman, J.B. Hirschberg, and M. Swerts, Predicting Automatic Speech Recognition Performance Using Prosodic Cues. *Proceedings of ANLP-NAACL*, pp. 218-225, 2000.

[5] C.E. Williams, K.N. Stevens, Emotions and speech: Someacoustical correlates. *J. Acoust. Soc. Amer*. 52, pp 1238-1250, 1972.

[6] S.T. Polzin, and A. Waibel, Pronunciation variations in emotional speech. In H. Strik, J. M. Kessens & M. Wester (Eds.) *Modeling Pronunciation Variation for Automatic Speech Recognition*. Proc. of the ESCA Workshop, 1998, pp. 103-108.

[7] S.J. Young, Large Vocabulary Continuous Speech Recognition.*IEEE Signal Processing Magazine* 13, (5), pp 45-57, 1996.

[8] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie, and C. Cox, "ASR for emotional speech: clarifying the issues and enhancing performance", Neural Networks Elsevier Publications, Volume 18, Issue 4, pp 437- 444, 2005.

[9] C. Whissell, "The dictionary of affect in language". In R. Plutchnik & H. Kellerman (Eds.) Emotion: Theory and research. New York, Harcourt Brace, pp. 113-131, 1989.
[10] ERMIS FP5 IST Project http://manolito.image.ece.ntua.gr/ermis/
[11] EC HUMAINE project (http://www.emotion-research.net).
[12] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, M.Schröder, 'Feeltrace': An instrument for recording perceived emotion in real time. In E. Douglas-Cowie, R. Cowie & M. Schröder (Eds.) *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, Belfast, pp.19-24, 2000.
[13] E. Douglas-Cowie, et al. Multimodal data in action and interaction:a library of recordings and labelling schemes HUMAINE report D5d http://emotion-research.net/deliverables/ 2003.