14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

RECOGNISING VERBAL CONTENT OF EMOTIONALLY COLOURED SPEECH

Theologos Athanaselis, Stelios Bakamidis, and Ioannis Dologlou

Institute for Language and Speech Processing
Department of Speech Technology, Artemidos 6 and Epidavrou, GR-15125, Maroussi, Greece
phone: + (30) 2106875416, fax: + (30) 2106854270, email: tathana@ilsp.gr
web: www.ilsp.gr

ABSTRACT

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. 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.

1. INTRODUCTION

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.

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 and a description of each component follow in section 2. The 3rd section describes the performance of the system. The enhanced language model and results of using it are discussed in section 4 and concluding remarks are made in section 5.

2. SPEECH RECOGNITION SYSTEM

2.1 System's overview

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 $\mathrm{Y} = y_1, y_2, ..., y_n$, 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 $\mathrm{P}(W \mid \mathrm{Y})$ into two components, that is,

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

The prior probability $\mathrm{P}(W)$ is determined directly from the language model. The likelihood of the acoustic data $\mathrm{P}(Y \mid W)$ 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.

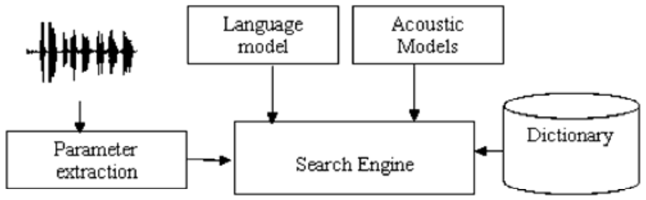

Figure 1 – The architecture of the speech recognition engine.

2.2 Description of speech recognition components

The prime function of the parameter extraction module is to divide the input speech into blocks and for each block to derive a smoothed spectral estimate. The spacing between blocks is typically 10 msecs and blocks are normally overlapped to give a longer analysis window of typically 25 msecs. A Hamming window weighting is applied to each block and Mel-Frequency Cepstral Coefficients (MFCCs) are used to model its spectral characteristics.

The purpose of the acoustic models is to provide a method of calculating the likelihood of any vector sequence Y given a word *w*. For a small vocabulary system, and digit recognition systems, we can have whole word models and achieve good performance. However for LVR systems this is impractical. In this case word sequences are decomposed into basic sounds called phonemes. Each individual phoneme is represented by an HMM. HMM phoneme models typically have three emitting states and left-to-right topology. For the English language, 45 phonemes are used to describe the pronunciation of all words. The corresponding HMMs were trained using the well known WSJCAM0 British English speech database comprising 8000 utterances (92 speakers, 90 utterances per speaker) [8].

The language model used by the LVR system is the standard statistical N-grams. 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:

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

The statistical language model of the LVR has been trained using BNC and consists of bigrams (N=2) and trigrams (N=3).

The basic recognition problem is to find the sequence of words that maximizes equation (1). LVR systems are quite complex requiring pruning of the search space. The search engine used here applies beam search and Viterbi decoding. The branching tree of HMM-state nodes are connected by state transitions and word-end nodes are connected by word transitions. Any path from the start node to an arbitrary point in the tree is denoted by a movable token placed in the node at the end of the path. The score of the token is the total log probability up to that point, and the history of the token records the sequence of word-end nodes that the token has passed through. For every time frame, the best score in any token is noted and any token that lies more than a beam-width below this best score is destroyed.

3. PERFORMANCE OF THE BASIC SYSTEM

3.1 Experimentation

The experimentation involves a large vocabulary continuous speech recogniser with basic language model and a test-set of sound files with emotional utterances derived by (Sensitive Artificial Listener) SAL software [9] developed as part of the ERMIS project (IST-2000-29319) [10].

3.2 SAL

SAL is mainly a software application designed to let a speaker work through various emotional states. 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 system contains four "personalities" that listen to the speaker and respond to what he/she says, based on the different emotional characteristics that each of the "personalities" possesses. The speaker controls the emotional tone of the interaction by choosing which "personality" they will interact with, while still being able to change the tone at any time by choosing a different personality to talk to.

What defines the personalities is that each one tries to steer the speaker towards a target state:

- Poppy is cheerful, and tries to steer the speaker towards a bright, positive state
- Spike is aggressive, and tries to steer the speaker towards an angry, confrontational attitude.
- Obadiah is gloomy, and tries to steer the speaker towards a passive, pessimistic outlook
- Prudence is matter-of-fact, and tries to steer the speaker towards a factual, matter-of fact attitude.

The core of the system consists of a range of responses designed to be applied when the speaker is in a particular emotional state, and has chosen a particular "personality" to talk [9].

3.3 Emotional Utterances

The sound files were drawn from a spontaneously emotional colored speech database [11], which is produced by people holding conversations with SAL. 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 (2 females (E, L) and 2 males (I, R)), is recorded with a sound quality that allows ASR. Speakers' emotional state is assessed by trained raters using the Feeltrace system [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 [12,13].

In order to evaluate speech recognition performance with respect to different emotional states (“Poppy”, “Prudence”, “Obadiah”, and “Spike”) 200 different sound files for each speaker were used (50 sound files per emotional state). Note that, using speech files with neutral style, speech recogniser performs robustly and recognition accuracy reaches up to 92%.

3.4 Experimental Results with basic language model

Figure 2 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 “Prudence”, 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 “Obadiah”, the lowest percentage of correctly recognised words is 17%, for speaker L, while the highest percentage of correctly recognised words is 55%, for speaker R. For all the speakers except speaker L, the speech recogniser performs better using “Obadiah” utterances.

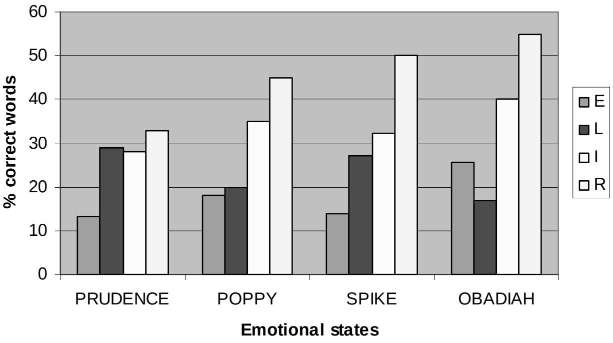

Figure 2 – The percentages of correctly recognised words for different emotional states and for different speakers, using a basic language model.

4. ENHANCING THE 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 [13] 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 [14]. 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 implies that each sentence of Whissell’s corpus will be appeared λ times on the emotional corpus and its value is adjusted experimentally to maximise recognition performance.

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

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

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 [13] has been experimentally established that best results are obtained for λ =10. The BNC contains about 6.25M sentences and 125K unique words. The Whissell’s corpus has 0.3M sentences. The emotional corpus has $9^{1/4}$M sentences; this figure is derived by Equation 3 [13].

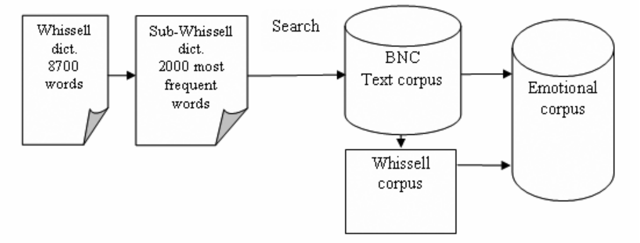
Whissell dict. 8700 words
Sub-Whissell dict. 2000 most frequent words
Search
BNC Text corpus
Emotional corpus
Whissell corpus

Figure 3 – The schematic view of text corpus enrichment using emotional sentences extracted from BNC corpus according to the sub-Whissell dictionary.

4.1 Experimental results with the enhanced language model

Figure 4 presents the percentage of correctly recognised words (y-axis) against the emotional states (x-axis) using the enhanced language model. The corresponding values for different speakers are depicted by the legend with variations in grey-scale color. In the case of emotional state “Poppy”, the lowest percentage of correctly recognised words is 25%, for speaker E, while the highest percentage of correctly recognised words is 65%, for speaker R. On the other hand, in the case of emotional state “Spike”, the lowest percentage of correctly recognised words is 23%, for speaker E, while the highest percentage of correctly recognised words is 70%, for speaker R.

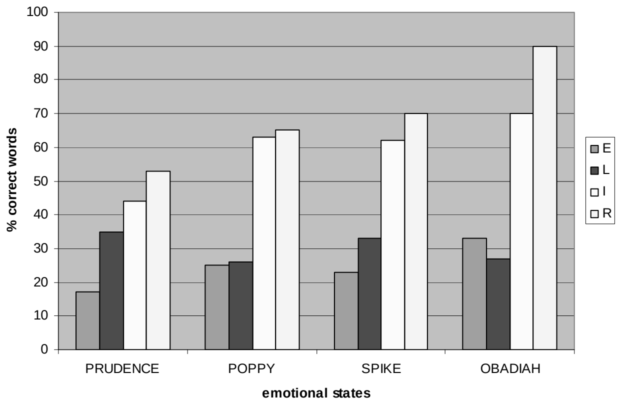

Figure 4 – The percentages of correctly recognised words for different emotional states and for different speakers using enhanced language model.

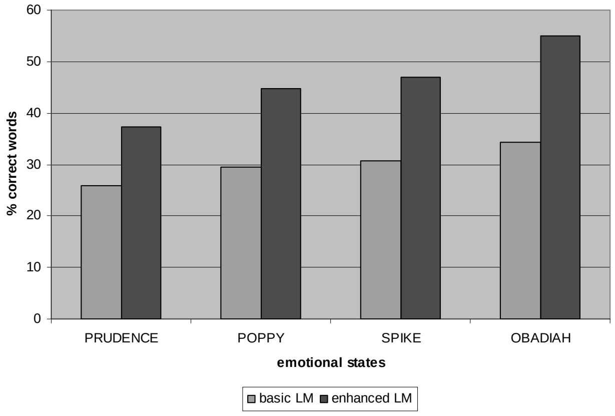

Figure 5 – The mean values of correctly recognised words for different emotional states with and without enhanced language model.

Figure 5 shows the improvement of mean value of correctly recognised words using enhanced language model. The mean value is computed by using the results for each speaker. The mean value of correctly recognised words for “Prudence” is 25,85% (basic language model), increasing to 37,25% (enhanced language model). For “Poppy”, is 29,45% increasing to 44,75%. For “Spike” is 30,8% increasing to 47%. For “Obadiah” is 34,4% increasing to 55%. These results imply that the largest improvement (20,6%) is related to “Obadiah” while the smallest improvement (11,4%) is observed with “Prudence”.

5. 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. Further testing should involve two key elements for improving speech recognition performance. The first refers to the design of specialized acoustic models and the second is to use a language model adapted to the linguistic structures that occur in emotional speech rather than analytic discourse. These challenges certainly invite further research.

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, 1999, pp 2079-2082.

[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*, 2000, pp. 218-225.

[5] C.E. Williams, K.N. Stevens, Emotions and speech: Some acoustical 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. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, 1995, pp 81–84.

[9] 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.

[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, 2000, pp.19-24.

[13] 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.

[14] 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.