

# Real-time Spoken Affect Classification and its Application in Call-Centres

Donn Morrison\*, Ruili Wang\*, Liyanage C. De Silva\*, W.L. Xu†

\*Institute of Information Sciences & Technology

†Institute of Technology & Engineering

Massey University, Palmerston North, New Zealand

donn.morrison.1@uni.massey.ac.nz, {r.wang, l.desilva, w.l.xu}@massey.ac.nz

**Abstract**—We propose a novel real-time affect classification system based on features extracted from the acoustic speech signal. The proposed system analyses the speech signal and provides a real-time classification of the speaker’s perceived affective state. A neural network is trained and tested using a database of 391 authentic emotional utterances from 11 speakers. Two emotions, anger and neutral, are considered. The system is designed to be speaker and text-independent and is to be deployed in a call-centre environment to assist in the handling of customer inquiries. We achieve a success rate of 80.1% accuracy in our preliminary results.

## I. INTRODUCTION

Along with the increased emergence of automated systems and computer interfaces in recent years, there has also been an increase in research on automatic recognition of human emotion or affect. It is now desirable to create systems that can recognise emotional variance in a user and respond appropriately [1].

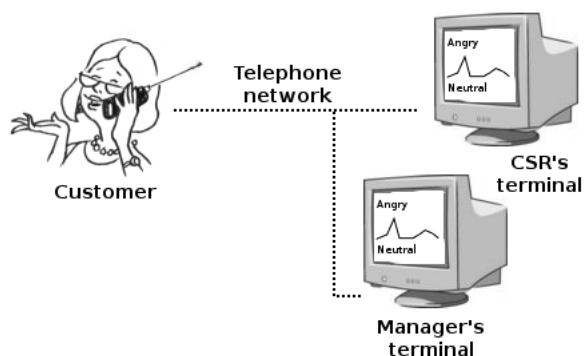


Fig. 1. Application of real-time affect recognition in a call-centre

Several studies have proposed systems for automatic recognition of human affect from speech [2]–[12]. In [11], a system was developed to automatically classify emotional speech using hidden Markov models based on features derived from log-frequency power coefficients (LFPC). A system using acoustic and language-based features and linear discriminant classifiers was developed in [5]. Decision trees and features based on acoustic and language information were used in the automatic classification system described in [7].

However, as far as the authors are aware, none of these studies address issues involved with the development of a real-

time system. We propose a novel real-time affect classification system with the aim of aiding customer support representatives and their managers in call-centres. Figure 1 illustrates an application of our system within a call-centre where a real-time affective assessment during a call would be available to both the customer service representative (CSR) and the manager.

Call-centres often have a difficult task of managing customer disputes. Ineffective resolution of these disputes can often lead to customer discontent, loss of business and in extreme cases, general customer unrest where a large amount of customers move to a competitor. It is therefore important for call-centres to take note of isolated disputes and effectively train service representatives to handle disputes in a way that keeps the customer satisfied.

In [12], a system was designed to monitor recorded customer messages and provide an emotional assessment for more effective call-back prioritisation. However, this system only provided post-call classification and was not designed for real-time support or monitoring. Our proposed system is different because it aims to provide a real-time assessment to aid in the handling of the customer while he or she is speaking. Early warning signs of customer frustration can be detected from pitch contour irregularities, short-time energy changes, and changes in the rate of speech [13], [14].

Apart from call-centres, there are many other areas where a real-time affect recognition system would be useful. In [1], several areas are identified where affect recognition can potentially aid those interacting with automated computer systems. An example is given how a teacher, human or computer, could tutor more effectively by accurately reading a student’s affective state. In [11], the authors discuss how stress affects the accuracy of automatic speech recognition systems. A real-time stress-detection module mounted on an automated speech recognition system could potentially lead to increased accuracy.

This paper is organised as follows. In Section II, we discuss our speech corpus and the benefits of using real-world speech data. In Section III, we discuss correlations between characteristics of the speech signal and affective states. Section IV provides an overview of the features correlating to certain affective states and the methods we have used for feature extraction. In Section V, we introduce a method of classification based on a neural network. Section VI lists our preliminary results and in Section VII, we present conclusions and avenues

for future work.

## II. SPEECH CORPUS

In the past, research in emotion or affect recognition was often marred by a lack of real-world speech samples [3], [4]. Due to ethical and moral issues it is often difficult to acquire a database of spontaneous emotional speech. Because of this, most studies have used actors to gather emotional speech [5], [6], [10], [15]. However, databases of acted speech do not accurately reflect real-world spontaneous dialogue. In spontaneous dialogue, affect is often subtly represented and difficult to detect, even for humans [11]. Other studies attempt to elicit emotions using elaborate scenarios such as WoZ (Wizard of Oz) [7], [8]. In these situations, data is collected from naive participants interacting with a malfunctioning system, causing frustration. Although this technique brings the data closer to the real-world, the participants are not in a real scenario where stress can be accurately modeled.

In contrast to most past research, we collected real-world affective speech data from a call-centre providing customer support for an electricity company. Customers telephoned with general queries, problems and comments and each call was handled by a customer service representative. The affective content of the data is mainly neutral, with the second largest subset representing angry callers. Worthy of note is the range of anger represented in the data. We have considered angry calls ranging from subtle frustration to hot anger. Because call-centres are most interested in detecting angry speech, this data suits our purpose.

The data was sampled at 22050 Hz (16 bit) and stored in MPEG Layer III format at a bit rate of 32 kilobits per second. Each conversation was manually divided into utterances and subsequently sorted depending on the perceived affective content. Calls were selected from a total of 11 speakers, 2 male and 9 female. The resulting dataset comprised of 190 angry utterances and 201 neutral utterances. In total there were 391 utterances.

The maximum utterance duration was 13.60 seconds, the minimum was 0.52 seconds, and the mean duration was 3.25 seconds. Some of the data contained a small amount of background noise; however, we feel this is particularly useful as it will help in the construction of a system that is robust.

## III. ACOUSTIC CORRELATES TO AFFECTIVE STATES

Pitch, energy, and speaking rate are widely observed to carry the most significant characteristics of affect in speech [2]–[12]. In [13], anger is described as having “an increase in mean pitch and mean intensity.” Downward slopes are noted on the pitch contour and there are “increases in high frequency energy.” From our research and in [1], increased pitch range is also apparent (see Figure 2). Neutral speech is shown to have a flatter pitch contour, with energy more equally dispersed in the spectrum.

Figure 2 shows the pitch contours of two example utterances from our speech corpus. It can be seen that the angry sample has downward slopes, concurring with [13], and a greater

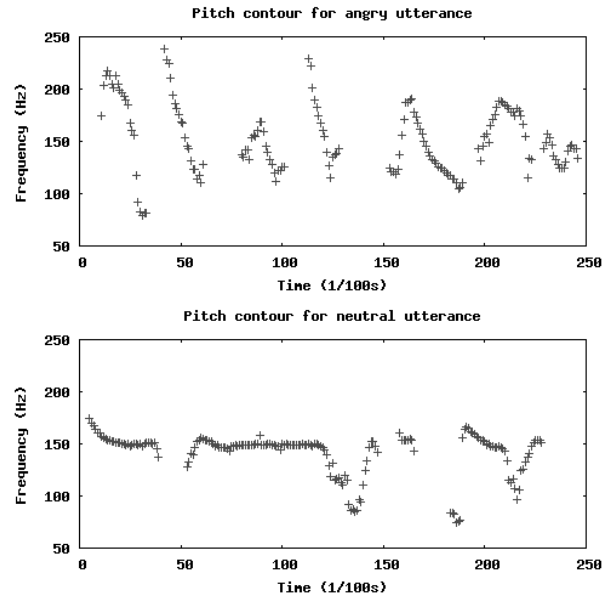


Fig. 2. Pitch contour for angry and neutral utterances

range. The neutral sample has a monotonous contour with a shallow range.

In [1], the changes in pitch for angry speech is said to be “abrupt on stressed syllables” and that the speaking rate of angry speech is typically faster than that of neutral speech. The same is found in [11], where the rate of speech is said to be high.

To summarise, when compared to neutral speech, anger has an increased pitch range and mean with the contour having abrupt downward slopes. The energy is much higher and is represented in the higher frequencies of the spectrum. The speaking rate is much faster. Based on these differences, we can build a neural network classifier that can be trained to recognise the patterns of angry and neutral speech. This will be discussed in Section V.

## IV. SIGNAL PROCESSING AND FEATURE EXTRACTION

### A. Endpoint Detection

During real-time processing, it is important for the system to be able to detect the endpoints of an utterance so that an assessment can be constructed. For an utterance the endpoints are defined as the beginning and the end. The endpoints can be detected by measuring the short-time energy on a frame-by-frame basis. When several frames have surpassed a threshold value for silence, signifying the beginning of an utterance, these frames can be buffered until the short-time energy falls below that threshold, signifying the end of the utterance. Figure 3 shows an example waveform with endpoints highlighted.

For this experimental setup, endpoint detection of the speech corpus was done by hand. Each database sample was segmented into utterances based on the above approach. In the implemented system, this step will be automated and will employ the aforementioned technique.

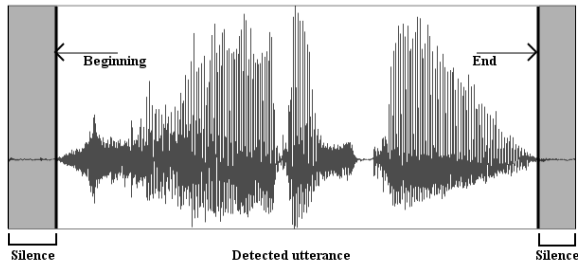


Fig. 3. Sample utterance with endpoints highlighted

TABLE I  
FEATURES EXTRACTED FROM EACH UTTERANCE

Feature group	Statistics
Fundamental frequency (F0)	maximum, minimum, mean, range, standard deviation, value at first voiced segment, value at last voiced segment
Formant frequencies (F1, F2, F3)	maximum, minimum, mean, range, standard deviation, value at first voiced segment, value at last voiced segment
Short-time energy	maximum, minimum, mean, range, standard deviation, value at first voiced segment, value at last voiced segment
Rhythm	speaking rate, average length of voiced segments, average length of unvoiced segments (pause)

### B. Features and Extraction Methods

Based on the acoustic correlates described in Section III, we have selected a set of 38 prosodic features as a starting point for describing the affective anger and neutral. These features were divided into four sets and are listed in Table I.

With the exception of those relating to rhythm, all features were calculated over the voiced segments of the sample. A frame is flagged as unvoiced if it has no value for the fundamental frequency.

Under the current implementation, features were extracted using a number of algorithms. The fundamental and formant frequencies were estimated using the RAPT algorithm (described in [16]) and linear predictive coding, respectively. A window size of 30 ms was used to estimate the fundamental frequency and is large enough for the estimation of a minimum pitch of 66.7 Hz, which is a reasonable lower bound for a male speaker. If the window is too short, lower pitch values (typically from male speakers) cannot be accurately estimated, however, if the window is too long, the resolution of the pitch contour is decreased [17]. The speaking rate was estimated by dividing the number of individual voiced segments by the total length of the utterance.

### C. Real-time Processing

The feature extraction process had to be very efficient because our time aim is real-time assessments. The continuous signal is processed and buffered into memory. When an endpoint is detected, the buffered speech is assumed to be a complete utterance and is sent to the feature extraction

TABLE II  
AVERAGE TIMES FOR FEATURE EXTRACTION COMPARED WITH THE AVERAGE LENGTH OF AN UTTERANCE IN THE DATABASE

Feature group	Average time (ms) <sup>1</sup>
Fundamental frequency (F0) contour	83.81
Formant frequency analysis	92.73
Short-time energy contour	30.49
Calculations <sup>2</sup>	76.27
Total	283.30
Utterance	3254.20

module. Here, features relevant to predetermined affective states are extracted from the utterance and are sent through the classifier, which provides an output in the form of an affective assessment. A block diagram is presented in Figure 4 to illustrate this process.

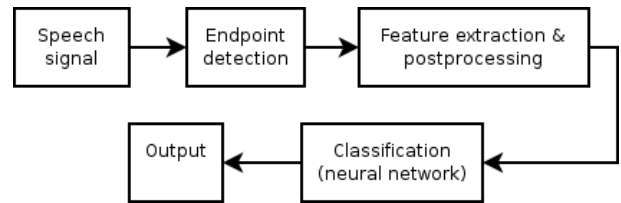


Fig. 4. Real-time vocal affect classification system

Table II lists calculation times for feature extraction. It is important to note that the total average time is substantially less than the total average time of an utterance. This shows that real-time utterance analysis is possible. In other words, if an utterance of length 3254.20 ms is buffered, an evaluation of that utterance will be available in 283.30 ms. This explains the possibility of real-time classification of the speech signal.

As the feature vectors are being calculated, new utterances are buffered and queued for processing. This concurrency ensures that the system does not begin to lag behind the speaker's rate of speech.

## V. CLASSIFICATION

Several past studies have used neural networks successfully for vocal affect recognition [8], [12]. Based on these studies, we have selected neural networks as the classifier for our study. Based on the size of the feature vector and the amount of available training data, we designed a 3-layer neural network consisting of 38 units in the input layer, a hidden layer consisting of 77 units, and an output layer with a single unit providing a binary classification of 1 or 0 (for neutral and angry, respectively). The speech corpus was divided into three groups. 70% of the data was used for training, 10% was used for validation during training, and the remaining 20% was used for testing. The network was trained using the RPROP learning algorithm described in [18], which is an improvement

<sup>1</sup>Benchmark times recorded on a Pentium 4 1.4 GHz CPU

<sup>2</sup>For example, maximum, minimum, mean, standard deviation, range (for pitch, energy, formants) and speaking rate

TABLE III  
CONFUSION MATRIX FOR ANGER AND NEUTRAL UTTERANCES

	Anger	Neutral
Anger	76.3%	23.7%
Neutral	16.1%	83.9%

over the standard back-propagation learning algorithm. This learning algorithm was also employed in [8].

To strengthen our classifier, we used  $v$ -fold cross-validation. With this method, the database is divided into equally sized subsets. With one subset left aside, the remaining are used to train the network. The network is then tested on the subset which was left out. This is repeated until every subset has been left out. This technique has been used previously in vocal affect recognition [8], [9], [12]. In our study, the database was divided into ten subsets ( $v = 10$ ). Each set was held out and used for testing while the other four were used for training and validation.

## VI. RESULTS

From the ten cross-validation sets, the minimum classification rate was 77.9% and the maximum classification rate was 87.3%. Table III shows the confusion matrix for angry and neutral speech utterances.

The overall classification rate was determined to be 80.1% for anger and neutral combined. It should be noted that a random classifier (coin toss) would yield a 50% accuracy rate for the two states. It should also be noted that the anger in our corpus ranges from subtle frustration to hot anger. By selecting only a subset of this wide range, we could increase the classification rate. However, one of our goals is to effectively predict the onset of hot anger by detecting a wider range of frustration.

To compare, we look at studies with similar characteristics to ours. In [8], a neural network was also used. The speech corpus was based on a Wizard of Oz scenario, so the authenticity of the speech was closer to the real-world than acted speech. They achieved a classification rate of approximately 60% for angry and neutral samples.

In [12], non-professional actors were used to gather speech data. Ensembles of neural networks proved to yield the highest classification rate of 75% for an emotion set of agitation and calm.

In [4], real-world data was used from interactions between users and an automated system. Using acoustic features, they achieved a success rate of 75-83% for negative and non-negative emotions.

The lack of standardised databases is probably the largest contributing factor to varying results between studies. Some studies use databases where differences between emotions are clear, while other databases contain speech that is easily misclassified even by human listeners. The choice of feature sets and feature selection methods also contribute to varying results.

## VII. CONCLUSIONS AND FUTURE WORK

We have proposed a novel real-time affect classification system with the aim of aiding customer service representatives in call-centres.

We succeeded in building a framework for real-time affect classification. This framework is speaker-independent because features were not selected depending on the speaker and nor was the classifier trained on a speaker-by-speaker basis. It is also text-independent because it does not use language features or contextual information. This allows the system to be used in areas without requiring speech recognition software.

We have satisfied a requirement for call-centres and have demonstrated the use of this system and the benefits it will have on the call-centre business.

During development of this experimental system, we looked at benchmark times for feature extraction algorithms. This is an important step and has been overlooked in all research on vocal affect classification to date as far as we are aware. In order for computer systems to be emotion-sensitive, it is imperative that they respond to input in a timely fashion.

We based our study on sets of features that have been discussed in past research. From our results in Section VI, we conclude that these features are a good starting point for finding more correlations between emotion and speech.

In the future, we plan to revisit several areas of this research. We must annotate and utilise more data from our speech corpus. We also hope to compare our spontaneous database with that of an acted database. This will help us to show that spontaneous emotion is much more subtle than that from actors and hence much more difficult to detect and predict.

## ACKNOWLEDGEMENTS

This study was funded by the Technology for Industry Fellowships (TIF), New Zealand. The authors are grateful for the use of the speech database provided by Mabix International.

## REFERENCES

- [1] R. W. Picard, *Affective Computing*. Cambridge, Massachusetts: The MIT Press, 1997.
- [2] A. Batliner, C. Hacker, S. Steidl, E. Noth, and J. Haas, "From emotion to interaction: Lessons from real human-machine dialogues," in *Affective Dialogue Systems*, E. Andre, L. Dybkjaer, W. Minker, and P. Heisterkamp, Eds. Berlin, Germany: Springer Verlag, 2004.
- [3] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth, "How to find trouble in communication," *Speech Communication*, vol. 40, pp. 117-143, 2003.
- [4] C. M. Lee and S. Narayanan, "Towards detecting emotions in spoken dialogs," pp. 1-30, January 19, 2004 In press, 2004.
- [5] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in the *International Conference on Spoken Language Processing (ICSLP 2004)*, Jeju Island, Korea, 2004.
- [6] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in the *International Conference on Spoken Language Processing (ICSLP 1996)*, Philadelphia, PA, 1996, pp. 1970-1973.
- [7] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in the *International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, Colorado, 2002.
- [8] R. Huber, A. Batliner, J. Buckow, E. Noth, V. Warnke, and H. Niemann, "Recognition of emotion in a realistic dialogue scenario," in the *International Conference on Spoken Language Processing (ICSLP 2000)*, vol. 1, Beijing, China, 2000, pp. 665-668.

- [9] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve, "Approaching automatic recognition of emotion from voice: A rough benchmark," in *the ISCA Workshop on Speech and Emotion*, Belfast, Northern Ireland, 2000, pp. 200–205.
- [10] T. S. Polzin and A. Waibel, "Emotion-sensitive human-computer interfaces," in *the ISCA Workshop on Speech and Emotion*, Belfast, Northern Ireland, 2000.
- [11] T. L. Nwe, S. W. Foo, and L. C. D. Silva, "Speech emotion recognition using hidden markov models," *Speech Communication*, vol. 41, pp. 603–623, 2003.
- [12] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Artificial Neural Networks in Engineering (ANNIE)*, St. Louis, Missouri, 1999, pp. 7–10.
- [13] K. R. Scherer, "Adding the affective dimension: A new look in speech analysis and synthesis," in *the International Conference on Spoken Language Processing (ICSLP 1996)*, Philadelphia, PA, 1996.
- [14] T. Johnstone and K. R. Scherer, "The effects of emotions on voice quality," in *the 14th International Congress of Phonetic Sciences*, San Francisco, CA, 2003.
- [15] F. Yu, E. Chang, Y.-Q. Xu, and H.-Y. Shum, "Emotion detection from speech to enrich multimedia content," in *the Second IEEE Pacific-Rim Conference on Multimedia*, Beijing, China, 2001.
- [16] D. Talkin, "A robust algorithm for pitch tracking (rapt)," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds. The Netherlands: Elsevier Science B.V., 1995, pp. 495–518.
- [17] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, New Jersey: Prentice-Hall, 1978.
- [18] M. Riedmiller and H. Braun, "A direct adaptive method for faster back-propagation learning: The rprop algorithm," in *the IEEE International Conference on Neural Networks*, San Francisco, CA, 1993, pp. 586–591.