

Spoken Affect Classification using Neural Networks

Donn Morrison, Ruili Wang, Liyanage C. De Silva
Institute of Information Sciences & Technology
Massey University, Palmerston North, New Zealand
donn.morrison.1@uni.massey.ac.nz, {r.wang, l.desilva}@massey.ac.nz

Abstract—This paper aims to build an affect recognition system by analysing acoustic speech signals. A database of 391 authentic emotional utterances was collected from 11 speakers. Two emotions, angry and neutral, were considered. Features relating to pitch, energy and rhythm were extracted and used as feature vectors for a neural network. Forward selection was employed to prune redundant and harmful inputs. Initial results show a classification rate of 86.1%.

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].

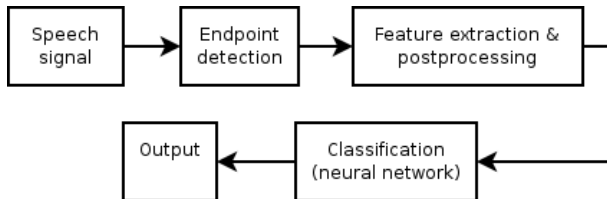


Fig. 1. Spoken affect classification system

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].

One application of automatic recognition of human affect is in a call-centre scenario. Call-centres often have a difficult task of managing customer disputes. Ineffective resolution of these disputes can 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.

Early warning signs of customer frustration can be detected from pitch contour irregularities, and short-time energy changes and changes in the rate of speech [13], [14]. Figure 1 shows a flow diagram of the proposed system.

Apart from call-centres, there are many other areas where an affect recognition system would be useful. In [1], several areas are identified where affect recognition can potentially aid those interacting with automated computer systems. For example, 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 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 discuss 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 speech depicting anger, this data suits our purpose.

TABLE I
SAMPLE UTTERANCES FROM THE DATABASE

Utterance	Affective state
This is insane.	Anger/frustration
It's on the account twice.	Anger/frustration
What do you mean?	Neutrality
This is under action - we will do something about it.	Anger/frustration
Right, we want to arrange to have our power put on.	Neutrality

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 or turns 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 data set comprised of 190 angry turns and 201 neutral turns. In total there were 391 turns.

The maximum utterance duration was 13.60 seconds, the minimum was 0.52 seconds, and the mean duration was 3.25 seconds. Table I lists some sample utterances and the associated affective state. 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 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 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.

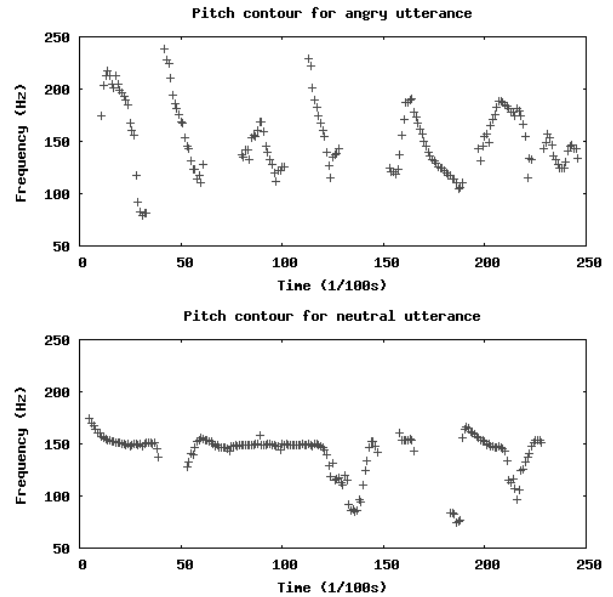


Fig. 2. Pitch contour for angry and neutral utterances

IV. FEATURE EXTRACTION

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 states anger and neutral. These features were divided into four sets and are listed in Table II.

TABLE II
FEATURES EXTRACTED FROM EACH UTTERANCE

Feature group	Statistics
Fundamental frequency (F0)	(1) mean, (2) minimum, (3) maximum, (4) standard deviation, (5) value at first voiced segment, (6) value at last voiced segment, (7) range
Formant frequencies (F1, F2, F3)	(8, 15, 22) mean, (9, 16, 23) minimum, (10, 17, 24) maximum, (11, 18, 25) standard deviation, (12, 19, 26) value at first voiced segment, (13, 20, 27) value at last voiced segment, (14, 21, 28) range
Short-time energy	(29) mean, (30) minimum, (31) maximum, (32) standard deviation, (33) value at first voiced segment, (34) value at last voiced segment, (35) range
Rhythm	(36) speaking rate, (37) average length of unvoiced segments (pause), (38) average length of voiced segments

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.

V. CLASSIFICATION

Several past studies have used neural networks successfully for vocal affect recognition [8], [12]. Hence, 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 with an output layer providing a binary classification of 1 or 0 (for neutral and angry, respectively).

Feature Selection

To help alleviate the *curse of dimensionality*, we employed *forward selection* to reduce unneeded features. Forward selection adds features to an existing set one at a time. The performance of the classifier is measured on this set at each stage. The resulting set consists of only those features which perform well together; all other irrelevant features are discarded.

TABLE III

ORIGINAL FEATURE SET (SET1) COMPARED WITH THE SET OBTAINED USING FORWARD SELECTION (SET2)

Label	Feature set
SET1	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38
SET2	30, 14, 8, 37, 35, 11, 15, 22, 2, 10, 1, 7, 36, 25, 13, 17, 28, 34, 9, 3

After forward selection, the most important features (30, 14, 8, and 37) were *minimum energy*, *range of F1 contour*, *mean of F1 contour*, and *average length of unvoiced segments (pause)*, respectively. This set consisted of 20 features and was labeled SET2.

Table III shows the features used in sets SET1 and SET2. Table IV shows that the performance of the neural network improved significantly through forward selection.

TABLE IV

COMPARISON OF CLASSIFICATION RESULTS ON FEATURE SETS

Classifier	SET1	SET2
Neural network	80.34%	86.12%

Neural Networks

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 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 nine were used for training and validation.

VI. RESULTS

Table V shows the confusion matrix for angry and neutral speech utterances.

TABLE V

CONFUSION MATRIX FOR ANGER AND NEUTRAL UTTERANCES

	Anger	Neutral
Anger	84.88%	15.12%
Neutral	12.65%	87.35%

The mean classification rate was determined to be 86.1% 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. While these results are not perfect, they are promising. 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. This higher success rate can be explained by the use of actors, as opposed to our use of real-world data.

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 presented a preliminary study that investigates the use of neural networks for classification of affective speech content. By employing forward selection on the feature vector, the performance of the network improved significantly.

In the future, we wish to increase the size of the database in order to train a more robust neural network. Other machine learning algorithms such as hidden Markov models or support vector machines will be studied and compared with the performance of the neural network.

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. Dybkaer, 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 backpropagation learning: The rprop algorithm," in *the IEEE International Conference on Neural Networks*, San Francisco, CA, 1993, pp. 586–591.