International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Identification of Emotions from Speech using Anchor Model

Surabhi G. Agrawal[1], Shabda Dongaonkar[2]

[1]Department of Computer Engineering, GHRCEM Wagholi, Pune University, India

[2]Professor Department of Information Technology, GHRCEM Wagholi, Pune University, India

Abstract: ***A technique for refining the anchor modelling system introduced enhanced emotion detection system from speech. Anchor representation was then put on the speaker detection issue. Identification mistake transaction productivity uncovered that the anchor displaying system missed the mark supply of a state-of the-craftsmanship GMM-UBM framework. It had been more observed that its computational efficiency was exceptional to that specific of the GMMUBM. Correlation of the anchor product and GMM-UBM programs for speaker indexing uncovered an indistinguishable trade-off between point of interest versus review productivity and computational efficiency. A cascaded speaker indexing system was arranged that uses the anchor item program as the first stage and the GMM-UBM as the second stage. In that configuration, the anchor system diminished the data pressing on the GMM-UBM while to some degree lessening effectiveness in working parts of low review. The impact of the cascaded project was to join the peculiarities of both projects at the inconvenience of some decrease in both computational productivity and accuracy of recognition. For extensive chronicles, the detection effectiveness of the anchor program and the absence of computational productivity of the GMM-UBM project can keep their application to speaker indexing. The cascading system may give a handy treatment to the speaker indexing application***.

Keywords: anchor modelling system, cascaded speaker, computational productivity, data pressing

1. Introduction

The anchor model was first introduced for speaker indexing in large audio databases [5] and then extended for speaker identification [6], speaker verification [7]. In this method, speaker identity is characterized by its relative position in an anchor space. This space is formed by a set of reference speaker models. Different metrics are used to compute the relative position of a given speaker with respect to the set of reference speakers such as Euclidean [5], angular [6], [7] or correlation [8] metrics. Several studies show that Euclidean distance achieves worse results compared to cosine distance [6], [7], [9]. In [7], a new qualitative measurement based on the rank metric was introduced. This new metric improves the performance compared to the quantitative distances but remains below the performance of the Gaussian Mixture Model-Universal Background Model (GMM-UBM) method. Mami and Charlet [6] have shown that anchor models perform better than GMMs when there is little amount of training data. In [10], the application of Linear Discriminate Analysis (LDA) post processing on coordinate vectors of the anchor space allows anchor models to outperform GMMs.

Furthermore, a system based on the combination of probabilistic and deterministic anchor model approaches has been proposed in [9] and achieves better results than the GMM-UBM based system. The probabilistic approach aims to model the intra speaker variability. Instead of representing the location of a speaker's utterances by only one point in the anchor model space, they are modelled using a normal distribution.

The anchor design system working as a characteristic extractor as opposed to as a combination method to recognize sensation from speech. So apply this technique to the particular job of knowing emotional speech of children interacting with a pet robot called AIBO.

The corresponding spontaneous emotional speech FAU AIBO Sensation Corpus, was presented and built publicly obtainable in Inter speech 2009, Challenge to provide community with a medium database deals with less prototypical and more spontaneous information to reveal more reasonable scenarios. Cepstral functions are produced to train GMM designs which can be applied as top conclusion of an anchor design system. In this study, shows that anchor designs is an efficient method to classify thoughts in the situation of highly unbalanced classes as is the event for the FAU AIBO Sensation Corpus.

Despite audio diarization and confirmation issues, anchor designs using easy range metrics such as the cosine full without the pre-processing stage achieve greater effects than GMM models. In addition, anchor designs perform a better than classifiers such as SVM- system based on back-end systems. There is an investigation the effectiveness of representing at prediction stage each sensation class by some consultant vectors on the other hand to an original vector.

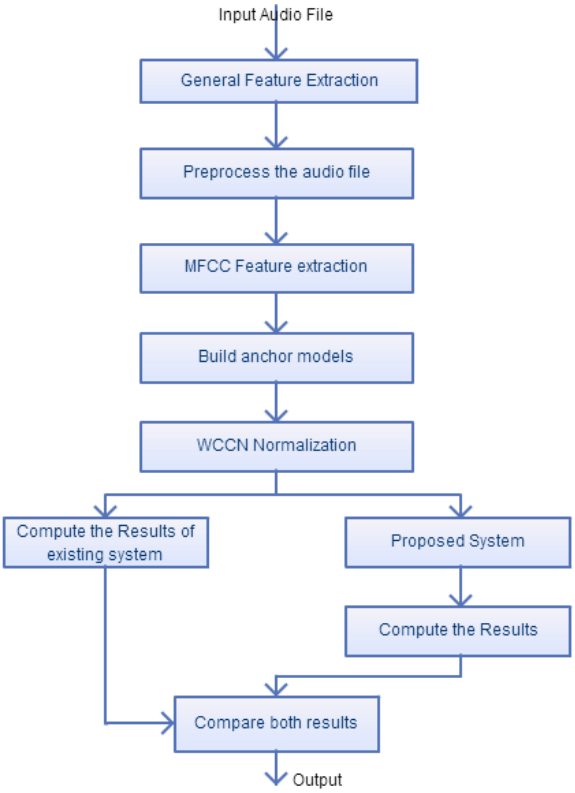

Figure 1: System architecture

2. Data and Features Description

2.1 Database

The Inter speech 2009 Emotion Challenge is another example, in which the FAU Aibo Emotional Speech Corpus [9], hereafter 'Aibo,' was used for training and testing the classifiers. This database consists of two separate parts, each recorded independently of the other at different schools with different participants. During the challenge, one part was used for training the classifiers, and the other for testing them. A chunk is an intermediate unit of analysis between the word and the turn manually defined based on syntactic-prosodic criteria. The average length of the chunk is about 1.7 s. The chunks are tagged into five emotion categories: Anger (A), Emphatic (E), Neutral (N), Positive (P, composed of baby talk and joyful), and Rest (R, consisting of emotions not belonging to the other categories such as bored, helpless, ...).

2.2 Feature Extraction

Extracting valuable features is another challenging task in the emotion recognition system. Mel frequency cepstral coefficients (MFCC) are one of the important features used in speech signal processing. Initially designed for speech recognition tasks they often give excellent performance in emotion detection tasks as well. The MFCC vector is formed of the first 12 coefficients including C0 (the zero cepstral coefficient as energy component) calculated at a rate of 10 ms using a 25 ms Hamming window. First and second derivatives are computed using a 5-frame window for each MFCC vector in order to compute the temporal characteristics. The Cepstral features are extracted using HTK toolkit [10]. The silences are removed from the audio files before the MFCC extraction.

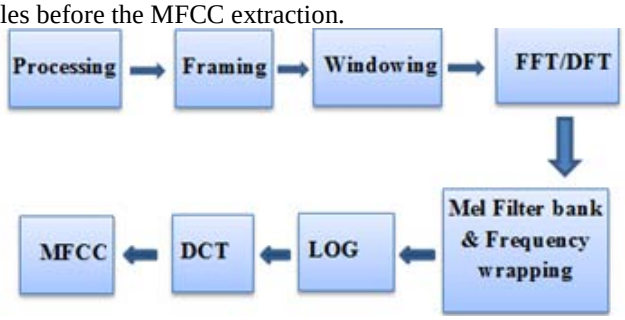

Figure 2: Process of MFCC calculation

3. Anchor Model

In associate anchor models system, an emotion class is characterized by its lives of similarity relative to alternative emotion categories. 3 steps characterize the look of associate anchor model system: building the anchor space, mapping the acoustic features onto the anchor space, and classifying test emotional speech.

3.1 Construction of Anchor Space

In the case of a pattern recognition drawback with unlimited range of categories like within the speaker verification task, like to seek out a group of speakers or virtual speakers (by cluster speakers) that is that the most representative of all speakers. Once the matter at hand involves a restricted range of categories, as for the feeling recognition task, the chance to model the whole set of feeling categories. Thus, during this kind of multiclass drawback, all categories have the advantage of being well depicted within the anchor area. Therefore, there was an illustration 2 main variation between the anchor models in speaker recognition and in feeling recognition. First, for speaker recognition the anchor area contains a high dimension, composed of many speaker models. For feeling recognition, the anchor area dimension is incredibly little thanks to the restricted range of feeling categories obtainable. Second, in speaker recognition, the speaker to characterize within the anchor area, throughout coaching or check stage doesn't typically belong to the set of anchor models. On the opposite facet in feeling recognition, the feeling appertains to the set of anchor models, owing that each one feeling category models are used as anchor models.

GMM is a generative model widely used in the field of speech processing. It is a probabilistic method that offers the advantage of adequately representing speech signal variability using a mixture of sufficient number of Gaussians. Given a GMM modelling a D-dimensional vector, the probability of observing a feature vector given the model is computed as follows:

$$P(\mathbf{x}|\lambda) = \sum_{i=1}^{m} w_i N(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

Where m, μk, and ∑_k correspond to the number of Gaussians, weight, mean vector, and diagonal covariance matrix of the kth Gaussian, respectively. GMM parameters are estimated using the maximum-likelihood (ML) approach based on the expectation maximization (EM) algorithm

3.2 Mapping

An anchor model system represents each emotion class relatively to the set of all other emotion classes. The set of these reference models is called anchor models. So call the new representation obtained by the projection of an utterance from the acoustic parameters space into the anchor model reference as Emotion Characterization Vector (ECV) by analogy to the speaker recognition terminology. The *C*-dimensional ECV space is generated by the models of the C-classes emotion recognition problem. An acoustic feature vector is mapped by computing its log-likelihood score for each emotion model.

$$\mathbf{L}(\mathbf{X}) = \begin{bmatrix} \log P(\mathbf{X}|\lambda_1) \\ \vdots \\ \log P(\mathbf{X}|\lambda_C) \end{bmatrix},$$

Where λ_i belongs to the set of class models {A, E, N, P, R} and $L(\mathbf{X})$ represents the ECV of **X**. At the test phase, each emotion class of the *C* classes is represented by a unique ECV vector obtained by computing the log-likelihood of all training data of that class against each of the *C* class models.

3.3 Classification

To classify a test speech, the distance between the ECV of a test data and those of each class representative is computed using either Euclidean or cosine metrics defined as:

- **Euclidean Matrix:**

$$d(\mathbf{L}_1, \mathbf{L}_2) = \sqrt{|\mathbf{L}_1 - \mathbf{L}_2|^2}$$

- **Cosine Matrix:**

$$d(\mathbf{L}_1, \mathbf{L}_2) = \frac{\langle \mathbf{L}_1, \mathbf{L}_2 \rangle}{\|\mathbf{L}_1\| \|\mathbf{L}_2\|},$$

Where <L1,L2> is the dot product of vector L1 & L2.

4. Conclusion

Anchor models, a similarity-based process, to fix the multiclass feeling acceptance problem. If WCCN normalization, Euclidean or cosine ranges can be indifferently applied as decision metric to significantly increase efficiency of the front-end system, specifically the GMM model. A relative obtain of 6.2 % is accomplished using Euclidean distance. There is indicated the number of the more complex and innovative classifiers applied as back-end methods can increase efficiency offered that the correct choosing or importance weighting process is used.

The most effective process, picked to overcome the issue of skewed class circulation, is classifier and characteristics dependent. Thus, by virtue of its algorithmic simplicity that does maybe not need any parameter focusing, its low time delivery complexity, and eventually its insensitivity toward unbalanced information, the anchor models system predicated on distance metrics represent a nice-looking alternative to boost on the efficiency of generative models such as for example GMM.

Reference

[1] YazidAttabi and Pierre Dumouchel," Anchor Models for Emotion Recognition from Speech" ,IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 4, NO. 3, JULY-SEPTEMBER 2013.
[2] StavrosNtalampiras and Nikos Fakotakis," Modelling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition", IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 3, NO. 1, JANUARY-MARCH 2012.
[3] Ali Hassan, Robert Damper," On Acoustic Emotion Recognition: Compensating for Covariate Shift",IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 21, NO. 7, JULY 2013.
[4] S. Mariooryad and C. Busso, "Compensating for Speaker or Lexical Variabilities in Speech for Emotion Recognition," In Press, Speech Comm., vol. 57, pp. 1-12, 2014.
[5] J. Arias, C. Busso, and N. Yoma, "Shape-Based Modeling of the Fundamental Frequency Contour for Emotion Detection in Speech," Computer Speech and Language, vol. 28, pp. 278-294, 2014.
[6] A. Metallinou, A. Katsamanis, and S. Narayanan, "A Hierarchical Framework for Modeling Multimodality and Emotional Evolution in Affective Dialogs," Proc. Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '12), pp. 2401-2404, Mar. 2012.
[7] D. Bone, M. Li, M. Black, and S. Narayanan, "Intoxicated Speech Detection: A Fusion Framework with Speaker-Normalized Hierarchical Functional and GMM Super vectors," Computer, Speech, and Language, Oct. 2012.
[8] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," Proc. 12th Ann. Conf. Int'l Speech Comm. Assoc. (Interspeech '11), pp. 3201-3204, Aug. 2011.
[9] D. Bone, M.P. Black, M. Li, A. Metallinou, S. Lee, and S. Narayanan, "Intoxicated Speech Detection by Fusion of Speaker Normalized Hierarchical Features and GMM Supervectors," Proc. 12th Ann. Conf. Int'l Speech Comm. Assoc. (Interspeech '11), pp. 3217- 3220, Aug. 2011.