International Journal of Scientific Research and Engineering Development-– Volume 3 Issue 3, May – June 2020

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Speech Activity Detection in Multi Party Meetings

[1]Thein HtayZaw, [2]Mie Mie Thaw
[1]Ph.D Student, University of Computer Studies,Mandalay (UCSM)
[2]Faculty of Computer Systems and Technologies, University of Computer Studies, Mandalay (UCSM)
theinhtayzaw@ucsm.edu.mm, miemiethaw@ucsm.edu.mm

--************************----------------------------------

Abstract:

This paper presents initial work towards the development of speech applications that involve multi-party conversation analysis. The aimof this work is to investigate the combination offeatures and Speech Activity Detection (SAD) classification approaches. The system conducts natural multi-talker conversational speech experiments in the Augmented Multiparty Interaction (AMI) meeting corpus to assess system performance. The proposed model based SAD system has been found to be comparatively better than other baseline SAD techniques for different types of meetings on frames that go as low as 10 milliseconds.

Keywords **—multi-party conversations, speech activity detection, classification approach**

--************************----------------------------------

I. INTRODUCTION

Speech Activity Detection (SAD) is an important task in most requests for speech processing. SAD system is the task of separating silence, non-speech frames from voice signals. The existence of non-speech frames has been found to considerably affect system performance. The detection of speech activity is also known as voice activity detection. However, as in speech processing terminology voice and speech are not same, the system call the task of identifying speech frames as SAD all through this work [1].

Energy-based SAD[4, 5] is by far the most popular SAD in the verification of speakers, possibly because of its uncomplicatedness. It determines the energy of each short-term frame and assumes, respectively, that the low- and high-energy frames correspond to nonspeech and expression. Standard SADs such as g729[6], ETSI advance front-end (AFE)[7] and statistical model SADs[8] have been developed for telecommunications desiderata and automatic speech recognition (ASR), namely low complexity and operation in real time. However there are applications, such as speaker diarization and identification for screening, indexing, or forensic use cases,that do not need real time processes.Recent research centred on energy-based features, probablyin combination with the zero-crossing rate (ZCR)[2], [3] and [19]. However, in the presence of additivenoise these features are highly impacted. Since the conditions for channel, noise and recording vary greatly from one meeting room to another, these systems typically do not make generalizations well to different recordings of meetings [9]. To classify the classes in unlabeled data, model-based detectors use pre-trained models on labeled data from speech and non-speech classes [10].The pre-trained models can also be adapted to the test data [11]. Gaussian mixture models are typically measured for each class and the detection using these models is based on Viterbi decoding.

Hybrid SAD approaches were proposed to address the issue of the generalization of pre-trained models to new conditions [12]. These methods combine both threshold-based energy detectors and model-based detectors for detecting speech / non-speech.

With regard to the state of the art, the key points of this research work are as follows: I proposing the use of robust source-related features for SAD purposes, ii) analysing the relative performance of energy-based and model-based SAD, iii) Examining best strategies to integrateinformation from various feature sets, iv) comparing the proposed SAD method with other algorithms on multi-party conversations The article is set out as follows. Section II describes the proposed features. The protocol used in our research is presented in Section III, and the results are discussed in Section IV. The paper is eventually concluded in Section V.

II. FEATURES

For practice a combination of complementary features is ideal because of the variety of speech characteristics [13]. Therefore, the system examine various characteristics of speech and evaluate to what extent the features reflect them.

A. Pitch

In several places the fundamental frequency of a sound, whose perception is called pitch, is of great importance. Effective pitch detection is important for areas such as speech coding, speech synthesis, and a more recent issue of recognition of speaker emotions [15]. Short-term autocorrelation (ACF) functions play a key role in speech processing, especially in pitch detection algorithms (PDA) and voicing dependent speech activity detection (SAD) functions [15]. A short-time autocorrelation function for a non-stationary signal, such as speech, which operates on short segments of the signal as:

$$R_x(m) = \frac{1}{N}\sum_{n=0}^{N'-1}[X(n+l)w(n)][X(n+l+m)w(n+n)]\ (1)$$

$$0 \le m \le m_0 \quad (2)$$

Where *w(n)* is a reasonable window for analysis, N is the length of the segment to be analyzed, N' is the number of signalsamples used in *R(m)* calculation, M_o is the number of autocorrelation points to be measured, and l is the index of the frame starting sample.For applications for pitch detection N' is generally set to the value given in (3):

$$N' = N - m \qquad (3)$$

B. Linear Prediction Error(LPE)

Many of the processing of speech, such as LPC techniques, is used for voice verification, speech synthesis, voice coding and recognition. LPC methods contain remarkably precise speech-parameter approximations. And that is quite ably what it does. And that is quite ably what it does. Linear prediction can be used to describe voice signal. Linear error in prediction indicates whether the signal is voice or not. As a linear combination of past samples, the fundamental principle of linear prediction is the present speech sample. The basic principle of linear prediction iscurrent speech sample can be closely estimated as a linear combination of past samples, as follows,

$$\hat{s}(n) = \sum_{k=1}^{p} a_k\, s(n-k)\ (4)$$

Where p represents the order of prediction, and where a k is the coefficients of linear prediction. The prediction error *e(n)* is the result of the difference between the actual sample s(n) and the predicted sample *s(n)*. One can calculate the prediction error *e(n)* as follows:

$$e(n) = s(n) - \hat{s}(n)\ (5)$$

For LP, the predictor coefficients ($a_k{}'s$) are calculated by minimizing the sum of square differences (over a finite interval) between the actual speaking samples and those linearly predicted. The LP coefficients minimizing error in prediction *e(n)*.

In prediction the cumulative error can be calculated as follows:

$$E = \sum_{n=\infty}^{\infty} e^2\,(n) \quad (6)$$

C. Spectral Flatness Measure

Spectral flatness[14] or tonality coefficient is the ratio of the geometrical mean to the power spectrum arithmetic mean. The arithmetic mean is the mean or sum of 'N' sequences while the geometric mean is the Nth root of their products. Consequently spectral flatness measure is given as:

$$flatness = \frac{exp\,(\frac{1}{N}\sum_{n=0}^{N-1} ln\, x(n))}{\frac{1}{N}\sum_{n=0}^{N-1} x(n)} \quad (7)$$

Where,*x(n)* is the magnitude of bin number n. Note that a single (or more) empty bin yields a flatness of 0, so this measure is most useful when bins are generally not empty.

D. MFCC

The human auditory system is more sensitive to variations in low-value frequencies than to high value ones. Therefore the frequency distances between consecutive mel bands in high frequencies are very high compared with the low frequency values. To convert the frequency in Hertz to mel is defined as follows;

$$m = 2595\, log_{10}(1 + \frac{f}{700}) \quad (8)$$

The implementation steps to calculate MFCCs are

- The signal isframed into short frames.
- Measure aperiodgram for every frame PowerSpectrum calculation.
- Apply the melfilterbankto the power spectra, in every filterthe energy is summed..
- Take the logarithm of all filterbank energies.
- Take the DCT of the log filterbank energies.
- Maintain DCT coefficientsfrom 2-13, delete the rest.

E. Delta and Delta Data

MFCCs itself provide the power spectral envelope for a single frame andare determined to model the dynamics of the MFCCs, delta (differential) coefficients as:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \quad (9)$$

Where d_t represents the delta coefficient vector of length N for the frame *t* and $c_t$represents the MFCC vector for frame *t*. Delta-Delta (Acceleration) coefficients are computed in the same way, but they are calculated from the deltas, not the static coefficients.

III. EXPERIMENTAL SETUP

For Most of the algorithms proposed for VAD can be divided into two processing stages:

- First, to acquire a representation that discriminates between speech and noise, features are extracted from the noisy speech signal.

- In the second level, a detection scheme is applied to the characteristics that result in the final decision.

A. Database Description

The aim of the system was to use a corpus containing agreat number of speakers and noisy conditions to train the technique proposed. The AMI corpus is composed of around 100 hours of transcripts of meetings recorded at various locations (Idiap, TNO, Edinbrough). The corpus includes meetings based on natural as well as scenario. The microphone array channel is distant recording, and therefore includes reverberation and ambient noise, in addition to the naturally occurring overlapping regions of speech. The system used microphone array channel and training, development and testing dataset comprises of 11, 10 and 12 meeting recordings, respectively in this work.

B. Energy Based Classification

Energy based classification used the combination of pitch and LPE features for energy based classification. Linear prediction order is 13. Linear Prediction Error normalized with respect to the energy of the signal. Normalized correlation function is used for pitch detection. Range(Bandedges) is [50, 400]. The analysis of the audio signals is carried out by a Hamming window with a frame length of K = 480 samples (30 ms), along with a hop length of R = 160 samples (10 ms). The system used maximum threshold of θ and γ to classify noise and silence interval from speech. Threshold θ is used for pitch and γ is used for linear predictive error. All experiments are based on the test set described in Section II.

C. *Model Based Classification*

The system selected for an ANN for its discriminating attributes, its capacity to modelnon-linear structures and the convenience of the posterior probabilities it produces for classification experiments. An artificial neural network (ANN) or neural networks (NNs) is a computational or mathematical model inspired by human nervous system structural and functional characteristics [16]. Multiple interconnected groups of artificial neurons form a neural network. These neurons are interconnected using a connectionist approach to computing some knowledge. ANNs are of an adaptive nature, meaning that they alter their structure based on the information that passes through the network (internal or external). Artificial neural networks consist of simpleelements, called nodes,that work in parallel.ANNs are of an adaptive nature,meaning that they change theirstructure based on the information that passes through the network (external or internal). Training shall be conducted until aparticular output for a corresponding input appears.Depending on the difference between target and output, the network is updated and stops whenthe difference between targets and output is zero or minimum; that is, output matches input[17].

The system used Neural Net Pattern Recognition, Neural Network Toolbox 11.1. It is multiple layer perceptron ANN. It involves input layer, hidden nodes, and output layers. Back propagation algorithm is used for training ANN classifier. 17MFCC, 17Delta, 17 DeltaDelta and 1SFM features are used to train ANN model.

D. *Hybrid Approach*

Hybrid approaches to SAD have been proposed to overcome the issue of generalizability of the pre-trained models to new conditions [18]. In this research work, Hybrid approach combines energy based classification and model based classification. Algorithm of Hybrid approach is as follows:

```
▪ Load classification output of energy based classification;
▪ Load classification output of model based classification;
▪ For i=1:size(framelength)
▪ if(energycfoutput(i)==0||modelcfoutput (i)==0)
▪ hybrid_cf_output (i)=0;
▪ else
▪ hybrid_cf_output (i)=1;
▪ end
▪ end
```

Where 0 represents noise and silence frame and 1 represent speech frame

IV. RESULT AND DISCUSSION

The performance of the proposed SAD was evaluated on the AMI dataset. SAD results are as shown in tables. These table showed maximum, average and maximum accuracy results of 6 audios in test set. .

TABLE I
SAD ERROR RATES OF ENERGY BASED CLASSIFICATION

Meeting ID	Miss Speech	False Alarm	SAD Error
ES2004a	6.7	25.01	31.71
ES2007d	5.18	18.05	23.23
ES2009b	5.15	12.68	17.84
ES2010b	3.17	22.67	25.84
ES2012d	3.65	21.00	24.66
EN2009b	1.07	15.70	16.78

TABLE III
SAD ERROR RATES OF MODEL BASED CLASSIFICATION

Meeting ID	Miss Speech	False Alarm	SAD Error
ES2004a	5.97	21.55	27.53
ES2007d	6.08	13.91	19,99
ES2009b	4.37	9.42	13.8

ES2010b	6.99	14.89	21.88
ES2012d	7.32	12.09	19.42
EN2009b	12.58	10.14	22.73

TABLE IIIII
SAD ERROR RATES OF HYBRID CLASSIFICATION

Meeting ID	Miss Speech	False Alarm	SAD Error
ES2004a	11.54	16.79	28.33
ES2007d	10.36	12.16	22.52
ES2009b	8.76	7.99	16.76
ES2010b	9.4	13.25	22.66
ES2012d	10.53	11.44	21.98
EN2009b	13.18	9.94	23.12

In the first part of experiments, the system investigated the combination of different feature sets (pitch, LPC) are used to classify the audio input. The results for model based experiments is shown in table I. As expected, accuracy drops dramatically with energy based classification.

The system improved in miss speech rate but degrade in false alarm rate and SAD error. These features are highly affected in the presence of additive noise.

Now turn attention to the model based VAD, in which different feature sets (17MFCC, 17 Delta, 17 DeltaDelta, 1 SFM) are combined and trained on ANN classifier. The results are displayed in Table II. A SAD-related problem that the proposed system does not yet address is the detection and discard of background speech frames. As with non-speak segments, feeding background speech into the ASR system will have a negative effect on its performance. It can be found that the proposed approach clearly outperforms other two approaches under all circumstances, often by a large decrease in the SAD mistake.

In the third part of experiments, the system investigated the combination of energy based classification method and model based classification method. SAD error is degrade than energy based method but increase than model base method.A grate issue is the classification of speech with voiced non speech sounds such as laughter or humming. These sounds can also be described as a particular speaker on theone hand, which can have a negative effect on the SAD mechanism.On the other hand, these events are often not included on the speech transcripts, mainly when they occur within the context ofanother speaker's voice, so those regions that are (in this case incorrectly) marked in the reference as non speech sound and noise intervals.It may result in a fairly high false alarm ratefrom a voice detector once again being tested on those results.In summary, the overall results indicate that data can be handledwith different channel type and conference center by the model-based SAD. To sum up, the overall results show that the model-based SAD can address data with various channel types and meeting rooms.

V. CONCLUSIONS

The paper proposed three classification methods for microphone array recordings in a meeting of scenario and non scenario.

A noise robust frontend for model based VAD was presented that extract four feature sets, each consist of a different attribute of speech in the audio signal. By training a classifier on the combination of the features, a high accuracy in robust speech detection was achieved, and this despite the low dimensionality of the features. Future work involves to investigate the robustness of alternative feature representations and the use of more advanced neural nets, e.g. long short term memory networks, to further increase the performance of the speech/non-speech classifier.

REFERENCES

[1] Sahidullah, M. and Saha, G., 2012. Comparison of speech activity detection techniques for speaker recognition. arXiv preprint arXiv:1210.0297. J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[2] Lamel, L., Rabiner, L., Rosenberg, A. and Wilpon, J., 1981. An improved endpoint detector for isolated word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 29(4), pp.777-785.

[3] Kotnik, B., Kacic, Z. and Horvat, B., 2001. A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm. In Seventh European Conference on Speech Communication and Technology.

[4] *Kinnunen, T. and Li, H., 2010. An overview of text-independent speaker recognition: From features to supervectors. Speech communication, 52(1), pp.12-40.* "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.

[5] Sun, H., Ma, B. and Li, H., 2010, December. Frame selection of interview channel for NIST speaker recognition evaluation. In 2010 7th International Symposium on Chinese Spoken Language Processing (pp. 305-308). IEEE.

[6] Benyassine, A., Shlomot, E., Su, H.Y., Massaloux, D., Lamblin, C. and Petit, J.P., 1997. ITU-T Recommendation G. 729 Annex B: a silence compression scheme for use with G. 729 optimized for V. 70 digital simultaneous voice and data applications.

[7] *Doc, E.S., 2002. Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms. ETSI ES, 202(050), p.v1.*

[8] Sohn, J., Kim, N.S. and Sung, W., 1999. A statistical model-based voice activity detection. IEEE signal processing letters, 6(1), pp.1-3.

[9] Van Leeuwen, D.A. and Huijbregts, M., 2006, May. The AMI speaker diarization system for NIST RT06s meeting data. In International Workshop on Machine Learning for Multimodal Interaction (pp. 371-384). Springer, Berlin, Heidelberg.

[10] Zhu, X., Barras, C., Lamel, L. and Gauvain, J.L., 2007. Multi-stage speaker diarization for conference and lecture meetings. In Multimodal technologies for perception of humans (pp. 533-542). Springer, Berlin, Heidelberg.

[11] Huijbregts, M. and De Jong, F., 2011. Robust speech/non-speech classification in heterogeneous multimedia content. Speech Communication, 53(2), pp.143-153.

[12] Sun, H., Nwe, T.L., Ma, B. and Li, H., 2009. Speaker diarization for meeting room audio. In Tenth Annual Conference of the International Speech Communication Association.

[13] Van Segbroeck, M., Tsiartas, A. and Narayanan, S., 2013, August. A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice. In INTERSPEECH (pp. 704-708).

[14] A. Gray and J. Markel, "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis," IEEE Trans. Acoust. Speech Signal Process., vol. 22, no. 3, pp. 207–217, Jun. 1974, doi: 10.1109/TASSP.1974.1162572.

[15] Tan, L. and Karnjanadecha, M., 2003, September. Pitch detection algorithm: autocorrelation method and AMDF. In Proceedings of the 3rd international symposium on communications and information technology (Vol. 2, pp. 551-556).

[16] Yegnanarayana, B., 2009. Artificial neural networks. PHI Learning Pvt. Ltd.

[17] Dietterich, T.G., 2000, June. Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Springer, Berlin, Heidelberg.

[18] Sun, H., Nwe, T.L., Ma, B. and Li, H., 2009. Speaker diarization for meeting room audio. In Tenth Annual Conference of the International Speech Communication Association.

[19] Zaw, T.H. and War, N., 2017, December. The combination of spectral entropy, zero crossing rate, short time energy and linear prediction error for voice activity detection. In 2017 20th International Conference of Computer and Information Technology (ICCIT) (pp. 1-5). IEEE.