

A CNN based Speaker Recognition System using an Alternate Bone Microphone



Khadar Nawas K, A Nayeemulla Khan

Abstract: State-of-art speaker recognition system uses acoustic microphone speech to identify/verify a speaker. The multimodal speaker recognition system includes modality of input data recorded using sources like acoustics mic,array mic ,throat mic, bone mic and video recorder. In this paper we implemented a multi-modal speaker identification system with three modality of speech as input, recorded from different microphones like air mic, throat mic and bone mic . we propose and claim an alternate way of recording the bone speech using a throat microphone and the results of a implemented speaker recognition using CNN and spectrogram is presented. The obtained results supports our claim to use the throat microphone as suitable mic to record the bone conducted speech and the accuracy of the speaker recognition system with signal speech recorded from air microphone get improved about 10% after including the other modality of speech like throat and bone speech along with the air conducted speech.

Keywords : Throat Speech,Bone Speech,Speaker Identification,CNN,Multi-modal Speaker Recognition.

I.INTRODUCTION

Automatic speaker recognition is a way in which the machines are used to identify/recognize the speaking person using the speech information.ASR has been a research interest for many decades; the transition of the technologies used in ASR is the interesting key factor to make the research challenging one. The challenges includes in the feature extraction techniques, speaker modeling and in the decision making techniques. The features depict the identity of the speaking person and the modeling the features involves the representation of the speaker and these models are used to identify/recognize the speaker. The pipeline of the ASR system involves Speech data collection, feature extraction , model training ,model testing and the evaluation as shown below Fig: 1. The performance of the ASR depends on techniques and technologies used in each step in the pipeline. The quality of the speech depends on recording device and the ambiance of the recording environments sound vibrations in the air ,whereas the throat pickups the sound vibrations near the vocal chords and the bone mic pickups the sound vibrations from the bones like skull. The AM signals contain the environmental back ground noise.

The TM and BC signals are in-contact with skin/surface, that are void from the back ground noise.

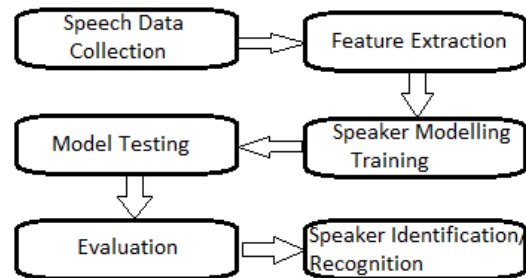


Fig : 1 ASR System Pipeline

sound vibrations in the air ,whereas the throat pickups the sound vibrations near the vocal chords and the bone mic pickups the sound vibrations from the bones like skull. The AM signals contain the environmental back ground noise. The TM and BC signals are in-contact with skin/surface, that are void from the back ground noise.

Air Microphone (AM)

The condenser microphone's speech is commonly used in speech processing studies. These data are referred as Air-conduction speech, a condenser mic capture the vibrations through the air medium and convert them to speech signals. The AM speech is affected by the background noise. The intelligibility of the AM speech signal get affected the background noise but the AM speech contains all the information from the higher to the lower frequencies.

Throat Microphone (TM)

The throat mic uses the piezoelectric transducer to sense the vocal cord vibration that is positioned near the larynx in contact with the skin of the throat. It collects the speech signals transferred by the sound vibrations along with the larynx tone. Because of its skin contact, it is less prone to the environment blare compared to the conventional microphone that senses the differences in air pressure and hence the environment noise gets captured. The speech of the throat microphone has less intelligibility due to filtering of the higher frequency by the skin and muscles at the larynx region, though it has speech signal with the speaker's characteristic features. The spectral features of some sound units differ from the normal microphone speech's sound units. There exits few distinctive spectral features in the TM speech compared to the AM speech. The presence of such spectral characteristics in the TM speech could be used to construct a speaker recognition system [1]. In the TM and AM voice, the spectral characteristics of certain sounds emerge to be complimenting one another by nature.

Revised Manuscript Received on December 30, 2019.

* Correspondence Author

Khadar Nawas K*, SCSE, Vellore Institute of Technology, Chennai, India.

A Nayeemulla Khan, SCSE, Vellore Institute of Technology, Chennai, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

A CNN based Speaker Recognition System using an Alternate Bone Microphone

The existence of such complimentary speaker specific spectral features of both voice signals results in increased efficiency of speaker recognition systems.

Bone Speech The technology in the bone conduction had a fold growth in recent decades. In both civil and military communications structures, bone conduction (BC) vibrators and BC touch microphones have become available as a radio conversation interfaces. BC speech has main advantages over voice communication interfaces with air conduction (AC), including minimal noise caused by the adverse environmental condition. The bone conduction path is shown in Fig 2. The BC microphone pickups the vibration of the skin and the bone, and it believe that the speech of bone-conduction involves private characteristics of speaker from those of the air conducted speech and throat speech. The BM speech lacks the information at higher frequencies as the TM speech due the filtration higher frequency component by the skin and the muscle along the sound transmission path . The background noise did not affect these voice information because they were recorded over the skin surface near the skull bone[2][3][4].

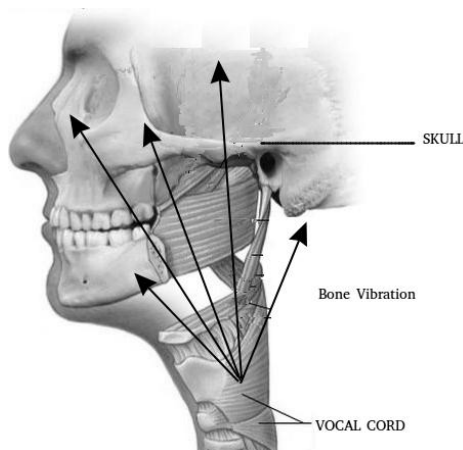




Fig: 2 Transmission of speech through bones and muscles

Bone Speech recorded using TM

In this work the throat mic is used to record the speech from the subject speaker. The need of using Throat mic is for its easy availability and the cost, even readily available in the local online commercial ecommerce websites. Since the Bone conduction mic cost more and it is not available in country like India. The data bone speech data set is rare and it is not publicly available for the researches to work in the bone microphone. We tried recording the bone vibration and able to record the speech. We also analysis the bone speech of both the speech recorded using the throat microphone and the original bone microphone Temco HG-17. The Spectral analysis of different speech like air, throat and bone is given in the figure. The spectral analysis of the bone speech from the Throat mic and Temco Hg-17 posses the same frequency elements with varying intensities. This evidence the speech information is recorded by the throat mic from the skull bone. The placements of the bone mic near the speaker head was studied and their speech intelligibility was reported high near the chin region, next collar bone, vertex and next condyle [3][4]. We used the condyle location near the ear as shown in figure No: 4 to record the speech from speaker head[7]. The hardware technical specification of both the throat mic and the original bone mic is presented in the TableNo:1. From the table the working principle of the original bone conduction microphone and the throat

microphone are different hence there will be difference in the outcome of the signal, still both the microphone sense the vibration, by taking this a advantage we used the throat microphone as a recording device for the bone speech.

Table:1 Specification of Throat and Bone conduction Microphone

| Microphone Type | Throat Microphone | Bone Conduction Microphone(Temco HG-17) |
|--------------------|---|---|
| Conduction Type | Vibration sound conversion Microphone | Vibration sound conversion Microphone |
| Frequency Range | 300Hz to 3400Hz | 200Hz to 4000Hz |
| Physical structure |  |  |
| Condenser type | Piezoelectric transducer | Electrets condenser |
| Wearable location | Neck | Head |

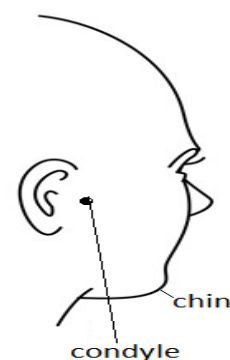
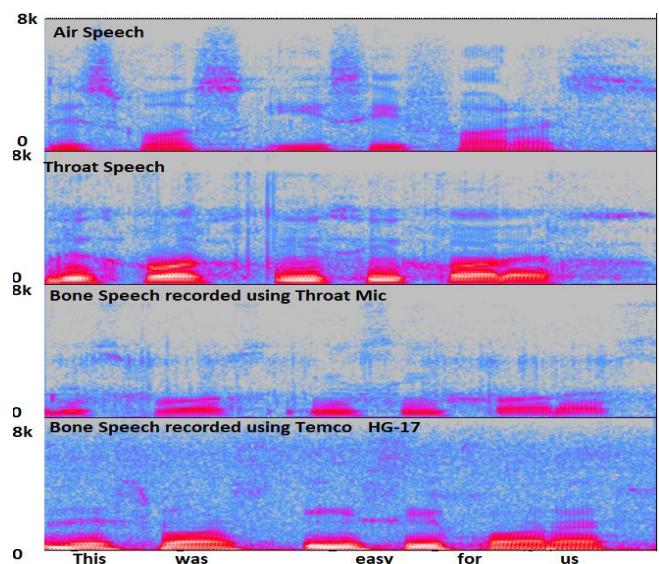


Fig: 4 Throat Microphone position

II. DATA COLLECTION AND EXPERIMENTAL SETUP

This study's database includes recordings made under laboratory circumstances where the ambient noise caused by the fan, door and other human noises are possible. 20 Volunteer's speech is obtained using the throat, air microphones and bone microphone asynchronously. This research uses text-dependent speech, Thirty TIMIT sentences are used as utterance for recording [5][6]. The below in Table no 2. The recordings are performed in distinct sessions to train and test the speaker models. 90 utterances is recorded from each speaker with three microphone. For each microphone 30 utterance from each speaker. The recorded speech possess session variability and channel variability. Two different laptops namely lenovo E41,Dell are used to record speech from 6 speakers and 14 speakers respectively. Total 1800 utterance are used to test and train the speaker models. The speech recorded from AM is not a clean speech which possess the lab environment noise such as fan noise, door noise etc. The TM and BM speech are clean with back ground and possess little noise due to sensor disturbance due head motion while recording. Gray scale spectrograms are converted and stored as PNG image format training and testing the CNN.

Table 2 TIMIT sentences.

| Name | TIMIT sentence | Name | TIMIT sentence |
|------|--|------|--|
| sa1 | She had your dark suit in greasy wash water all year | sx16 | A roll of wire lay near the wall. |
| sa2 | Don't ask me to carry an oily rag like that | sx17 | Carl lives in a lively home |
| sx3 | This was easy for us | sx18 | Alimony harms a divorced man's wealth. |
| sx4 | Jane may earn more money by working hard. | sx19 | Aluminum silverware can often be flimsy. |
| sx5 | She is thinner than I am | sx20 | She wore warm, fleecy, woolen overalls. |
| sx6 | Bright sunshine shimmers on the ocean | sx21 | Alfalfa is healthy for you |
| sx7 | Nothing is as offensive as innocence. | sx22 | When all else fails, use force |
| sx8 | Why yell or worry over silly items? | sx23 | Those musicians harmonize marvelously. |
| sx9 | Where were you while we were away? | sx24 | Although always alone, we survive. |
| sx10 | Are your grades higher or lower than Nancy's? | sx25 | Only lawyers love millionaires. |
| sx11 | He will allow a rare lie. | sx26 | Most young rise early every morning |
| sx12 | Will Robin wear a yellow lily? | sx27 | Did dad do academic bidding? |
| sx13 | Swing your arm as high as you can | sx28 | Beg that guard for one gallon of gas |
| sx14 | Before Thursday's exam, review every formula. | sx29 | Help Greg to pick a peck of potatoes. |
| sx15 | The museum hires musicians every evening | sx30 | Get a calico cat to keep |

III. EXPERIMENTS AND RESULTS

The spectrogram is a form a representation of the speech signals in the frequencies domain as images. The spectrogram represents the frequencies range in y axis, intensity of the signal as shown in gray scale more the gray color the intensities is high and the timing information in x axis[8]. It consist of temporal information as well as spatial information of the signal. This type of gray scaled spectrogram images generated from the speech signals are used to perform the experiments. In recent decades Convolution neural network(CNN) has been evolved as major computer vision tool to solve many real-time complex application like image classification, object detection ,image enhancement and video analytics. In this work, Alex Net and googleNet CNN architectures are used in training and testing the speaker identification task. Generally CNN requires a huge volume of data and high computation resource for learning/training task. Whereas, the available limited data is not sufficient to train a CNN. Hence, transfer learning technique is adapted to over the limitation for the use of CNN. The existing Alexnet[9] and GoogleNet[10] are pre-trained networks, trained with imagenet data set. Though these networks are pertained ,it is required in our work to retrain the CNN with our dataset. 15% of dataset is used for testing,10 % of dataset is used for validation and remaining dataset is used to retrain the CNN.In four different modals the speaker identification experiments are carried as in following categories :

Using the Air conducted speech, Throat Speech and bone speech separate speaker identification experiments and by combining all the three speech the experiments are conducted and the obtained results are in Table 3 & 4.

Table:3 AlexNet Result

| Type of Speech | Accuracy(%) | |
|----------------------------------|-------------|------|
| | Trainings | Test |
| Air Speech | 73 | 73 |
| Throat Speech | 88 | 93 |
| Bone Speech | 90 | 93 |
| Combined speech(Air+Throat) | 86.45 | 78 |
| Combined speech(Air+Bone) | 89 | 83 |
| Combined speech(Air+Throat+Bone) | 92 | 84 |

Table:4 Googl Net results

| Type of Speech | Accuracy (%) | |
|-----------------------------|--------------|------|
| | Trainings | Test |
| Air Speech | 82 | 71 |
| Throat Speech | 88 | 93 |
| Bone Speech | 90 | 87 |
| Combined speech(Air+Throat) | 93 | 76 |
| Combined Speech(Air+Bone) | 90 | 84 |



| | | |
|----------------------------------|----|----|
| Combined speech(Air+Throat+Bone) | 93 | 82 |
|----------------------------------|----|----|

IV. CONCLUSION

The necessity and need for such above experiments is to advance the accuracy of speaker identification system by using multi modal speech. The Bone microphones are used in research work conducted mostly by the military and the police organization of various countries. The commercial availability of the bone microphone is less possible than the commercial available throat microphone in the market. The publicly unavailability and lack in publicly available throat speech and bone speech dataset motivated the idea of using the throat mic as an alternate sensor to record the bone speech. From the table 3 &4 the identification accuracy of the air conducted speech is 73% and 71% in Alexnet and GoogleNet respectively .The less accuracy was resulted due to the quality of air speech. It is influenced by the environmental noise. Whereas the identification accuracies of the Throat speech and bone speech are higher than the air speech because theses speech is free from the background noise. When the all the three speech put to build the identification task the accuracy of the air speech is increased by 10%.From the above work it is concluded that the throat mic can be used as alternate microphone to record the bone speech.

REFERENCES

1. Mubeen, N., Shahina, a., Khan, a. N., & Vinoth, G. (2012). Combining spectral features of standard and throat microphones for speaker identification. International Conference on Recent Trends in Information Technology, ICRTIT 2012, 119–122.
2. McBride, M., Tran, P., Letowski, T., & Patrick, R. (2011). The effect of bone conduction microphone locations on speech intelligibility and sound quality. Applied Ergonomics, 42(3), 495–502.
3. Blue, M., McBride, M., Weatherless, R., & Letowski, T. (2013). Impact of a bone conduction communication channel on multichannel communication system effectiveness. Human Factors, 55(2), 346–355. <https>
4. McBride, M., Tran, P., Pollard, K. A., Letowski, T., & McMillan, G. P. (2015). Effects of Bone Vibrator Position on Auditory Spatial Perception Tasks. Human Factors, 57(8), 1443–1458.
5. Marx, M. A., Vinoth, G., Shahina, A., & Khan, A. N. (n.d.). Throat Microphone Speech Corpus for Speaker Recognition, 16–20.
6. Larcher, A., Lee, K., Ma, B., & Li, H. (2012). RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases. Interspeech, 2–5.
7. Blue, M., McBride, M., Weatherless, R., & Letowski, T. (2013). Impact of a bone conduction communication channel on multichannel communication system effectiveness. Human Factors, 55(2), 346–355.
8. Kekre, H. B., Kulkarnii, V., Gaikar, P., & Gupta, N. (2012). Speaker Identification using Spectrograms of Varying Frame Sizes. International Journal of Computer Applications, 50(20), 27–33.
9. Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton” ImageNet Classification with Deep Convolutional Neural Networks”.
10. Ren, J., Hu, Y., Tai, Y.-W., Wang, C., Xu, L., Sun, W., & Yan, Q. (2016). Look, Listen and Learn - A Multimodal LSTM for Speaker Identification, 3581–3587.