Unsupervised Speaker Recognition Based on Single Channel Speaker Count Estimation

K.Vasanthakumar
Department of CSE
Francis Xavier engineering college
kvasanthkrishnan@gmail.com

R.Vinothkumar
Department of CSE
Francis Xavier engineering college
kumarvinoth96rj@gmail.com

M.Mukeshkrishnan
Assistant Professor
Department of CSE
Francis Xavier engineering college

Abstract: Estimating the maximum number of concurrent speakers from single-channel mixtures is a challenging problem First step to address various audio-based tasks such as blind source separation, speaker diarization andaudio surveillance The convolutional recurrent neural networks outperform recurrent networks used in a previous study when adequate input features are used. The efficient performing method to several baselines, as well as the influence of gain variations, different datasets, and reverberation. This concept describes a system for unsupervised speaker recognition based on the piecewise-dependent-data (PDD) clusteringmethod.Thereare16conversationsof high-quality conversations between two and three participants, the estimation of thenumber of the participants was correct. In telephone- quality the results werepoorer.

Keywords: Competitive learning, segmentation, self-organizing maps(SOMs),speaker recognition, temporal data clustering, vector quantization(VQ).

I. INTRODUCTION

Speaker recognition describes a system for unsupervised speaker recognition (otherwise known as “speaker segmentation”), based on

the piecewise-dependent-data (PDD) clustering method. Most speaker recognition system problems have been solved by using supervised methods. A survey of issues and methods regarding supervised speaker recognition canbe found in . The training data for each speaker is given a priori and a model of each speaker is produced.Supervised methods have been applied for speaker identification and verification process, for example, for entering computers or security sites by vocalpasswords.

A more less common problem is unsupervised speaker recognition (speaker segmentation), in this case, no training set is given and the data is unlabeled. Since no labeled training dataset is available,theunsupervisedtrainingisperformed by initially clustering the data into different clusters where each cluster, is assumed to represents a different speaker. Unlike most clustering approaches where each vector is associated with a specific cluster, here a sequence of vectors has to be associated with a samecluster.

2. EXISTINGSYSTEM

Agoodworkingidolizationsystemwouldbeable to sufficiently address the speaker count estimation problem using this strategy. It appears to be a very complex problem to tackle when one is only interested in the number of concurrentspeakers.

The deep learning to an existing task, often is a matter of choosing a suitable network architecture. Typically an

architecture describes the overall structure of the network including (but not limited to) the type and number of layers in the network and how these layers are connected to each other. The complexity of a speaker recognition problem depends on the speaker

population size and the duration of the speaker speechsegment.supervisedspeakerrecognition problems depend on the following types one is text-dependent or another one is text- independent, on whether the set is closed or open, and on whether the problem is to identify or to verify thespeaker.

1 Disadvantages:

Existing system , speaker recognition problems depend on signal bandwidth (the telephoneline bandwidth), environmental noise, whether or not real time problem is solved , and the equipmentinuse,suchasthespeedandsolution of the sampler , the microphone type. And also theexistingsystemprovideslowlatencyandalso it requires high powerequalization.

3.PROPOSED SYSTEM

The proposed model uses a convolutional recurrent(CRNN) architecture, based on classification at the network's output . the proposed model and also suggests that for improvedbustnesstrainingwouldbenefitfroma largevarietyoftakingrates.TheproposedCRNN focuses on the temporal segmentation of phonemes. The proposed CRNN error bars show 95% confidenceintervals.

TThe Fig:1 graphs are represents the high quality conversation validityfunctions.

First one shows Twelve high-quality conversations between twospeakers.

Second one shows Five high-quality conversations between threespeakers

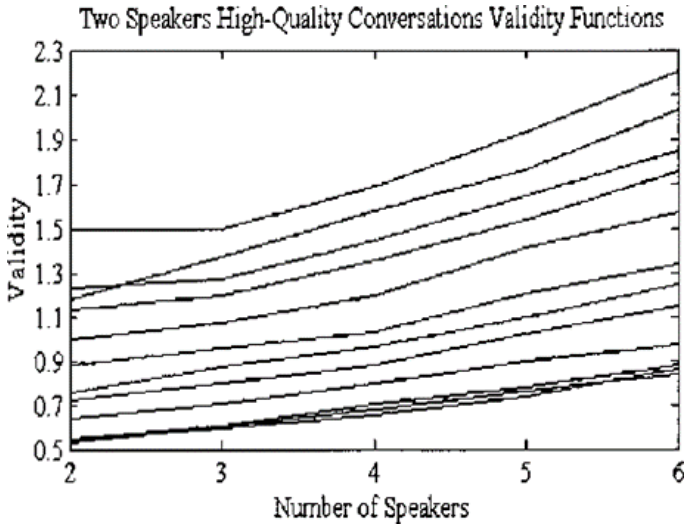
Two Speakers High-Quality Conversations Validity Functions
Validity
Number of Speakers
2.3
2.1
1.9
1.7
1.5
1.3
1.1
0.9
0.7
0.5
2
3
4
5
6

Fig:1

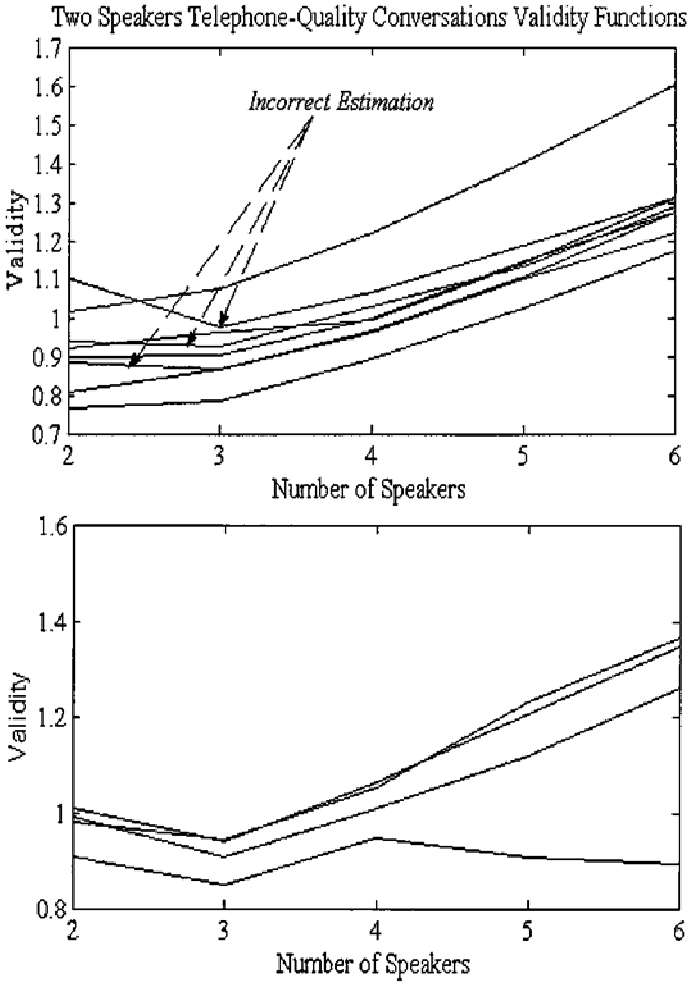
Two Speakers Telephone-Quality Conversations Validity Functions
Incorrect Estimation
Validity
Number of Speakers
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
2
3
4
5
6
Validity
Number of Speakers
1.6
1.4
1.2
1
0.8
2
3
4
5
6

Fig:2

Proposed system architecture

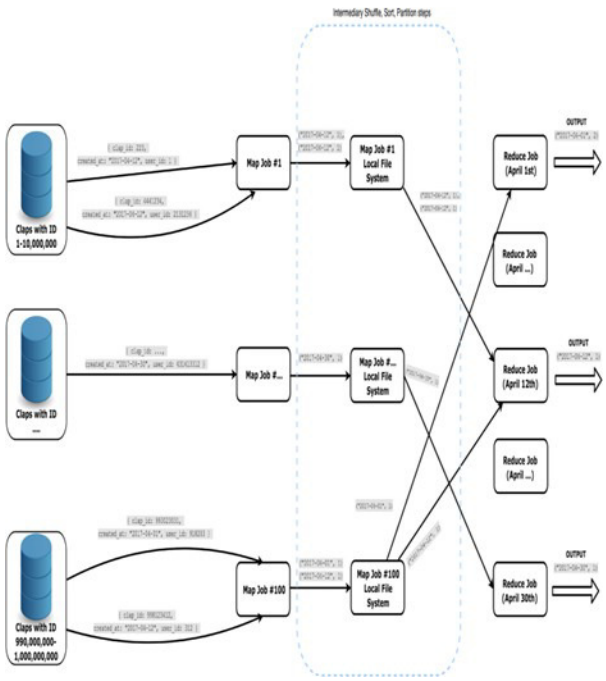

Fig:3 graphs are represents the telephone quality conversation validity Graphs

Eight telephone -quality conversations that converged well.

Four telephone -quality conversations that did not converge well.

Advantages:

System can work any kind of environment .The system also has minimum time complexity. The proposed system also get a different number of input frames ranging .And also the system supportvectormachine(SVM)witharadialbasis function (RBF)kernel.

4. CONCLUSION:

Speaker recognition system work a time-series clustering approach, based on an iterative process with competition between SOM models was investigated. Unsupervised speaker recognitiontaskseemstobeverydifficulttask.It is probably due to the fact that, to begin with, the amount of information about the speaker in speech signal is relatively low, as compared to theinformationonthemessage.Without*apriori* labeled information it is difficult to model the speakers.

The fact that in the general segmentation problem the number of speakers is unknown makes the problem extremely difficult. To achieve good clustering results we first had to determine the optimal size of the models to represent a speaker and the shortest segment length to derive sufficient statistic of the speaker.

Shorter segments enable better segmentation resolution. The experiments showed that SOM ofsize610issufficientforspeakermodeling.For short conversations (about 60 s for two-speaker conversations) segments of half second were needed. After all the segments length and models size were

defined, the effective algorithm was applied for high-quality conversations between two to five participants and for two-speakers telephone-quality conversations.

For two- and three-speakers high-quality data conversations and for two-speakers telephone-qualitydata,theresultswereusuallygood(more than 80% success). For four and five speakers, only one conversation of four speakers con- vergedcorrectly.

Tthe only conversation were the data was approximately homogeneously divided among the partici-pants, the average segment length was more than 3 s, there was almost no simultaneous data, and there was small amount of non speech data.

This shows that the algorithm is sensitive to the amount of data of every cluster, especiallywhen the data is over- lapping and, as well as to the amount of noise (nonspeech and simultaneous speechdata).Therefore,itmightbenecessaryto develop an effective speech/nonspeech and simultaneous speech detector. A validity criterion wassuggested.

Thevalidityestimationneveraffectedthequality oftheclusteringintwo-andthree-speakerhigh- quality conversation the estimation of the number of clusters was correct except for one three-speaker conversation. In telephone- quality eight out of 12 conversations were correctlyclustered.

Conversations contains five out of eight conversations the number of clusters was correctly estimated. In three conversations the number of speakers was estimated as three insteadoftwospeakers.Inordertocomparethe performance of the proposed approach to existing algorithms, a systematic review of the available literature was made. Four articles describing sev- eral algorithms appeared to be relevant forthe

comparison [12]–[15]. The algorithms cover differentvariationsofHMM.Duetothedifferent databases and the lack of information about the performed evaluations, only a very crude comparison could be made in the present study. It was found that in all these works all the conversations were at least 90 s. The results were similar to the reported here but in all the cases the algorithms were very sensitive to the initial conditions. The research of Cohen and Lapidus [14] and [15] was the only one done on the same telephone-quality database and the results weresimilar.

REFERENCES:

[1] S. Furui, “An overview of speaker recognition technology,” in*Proc. ESCA Workshop Automatic Speaker Recognition, Identification, Verifi- cation*, Apr. 1994, pp. 1–9.

[2] J. P. Campbell Jr.,“Speaker recognition: A tutorial,” *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sept. 1997.

[3] S. Furui, “Recent advances in speaker recognition,” *Pattern Recognition Lett.*, vol. 18, no. 9, pp. 859–872, Sept.1997.

[4] H. Gish, M.-H. Siu, and R. Rohlicek, "Segregation of speaker for speech recognition andspeaker identification," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 1991, pp. 873–876.

[5] M.-H. Siu, G. Yu, and H. Gish, "An unsupervised, sequential learning algorithm for the segmentation of speech waveform with multiple speakers," in *Proc. Int. Conf. Acoust. Speech, Signal Processing*, vol. 2, 1992, pp.189–192.

[6] T. Kohonen, *Self-Organization and Associative Memory*. Berlin, Germany SpringerVerlag,1989

[7] M. H. Kuhn, "Speaker recognition accounting for different voice con- ditions by unsupervised classification (cluster analysis)," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-10, pp. 54– 57, Jan.1980.

[8] A. Cohen and V.Lapidus,
"Unsupervised text independent speaker clas- sification," in *Proc. 18th Convention Elect. Electron. Eng.*
Israel, 1995, pp. 3.2.2 1–5.

[9] M. Sugiyama, J. Murakami, and H. Watanabe, "Speech segmentation and clustering based on speaker features," in *Proc. Int. Conf.Acoust. Speech, Signal Processing*, vol. 2, 1993, pp.395–398.