International Journal of Electrical and Computer Engineering (IJECE)
Vol. 6, No. 6, December 2016, pp. 2766~2772
ISSN: 2088-8708, DOI: 10.11591/ijece.v6i6.9844

On Usable Speech Detection by Linear Multi-Scale Decomposition for Speaker Identification

Wajdi Ghezaiel[1], Amel Ben Slimane[2], Ezzedine Ben Braiek[3]
[1]CEREP- ENSIT, University of Tunis, Tunis, Tunisia
[2,3]ENSI, University of Manouba, Manouba, Tunisia

Article Info

Article history:

Received Dec 31, 2015
Revised Mar 15, 2016
Accepted Mar 29, 2016

Keyword:

Co-channel speech
Discrete wavelet transform
Multi-scale decomposition
Speaker identification
Usable speech

ABSTRACT

Usable speech is a novel concept of processing co-channel speech data. It is proposed to extract minimally corrupted speech that is considered useful for various speech processing systems. In this paper, we are interested for co-channel speaker identification (SID). We employ a new proposed usable speech extraction method based on the pitch information obtained from linear multi-scale decomposition by discrete wavelet transform. The idea is to retain the speech segments that have only one pitch detected and remove the others. Detected Usable speech was used as input for speaker identification system. The system is evaluated on co-channel speech and results show a significant improvement across various target to Interferer Ratio (TIR) for speaker identification system.

Corresponding Author:

Wajdi Ghezaiel,
CEREP- ENSIT,
University of Tunis,
Tunis, Tunisia.
Email: wajdi.ghezaiel@gmail.com

1. INTRODUCTION

Speech can be distorted by many kinds of interferences. Interfering signal can be stationary or no stationary signal. Stationary noise can be dealt with denoising and noise reduction techniques; whereas non stationary noise is caused by another speech from a different speaker. Such interference is frequent and the corrupted speech is known as co-channel speech [1]. Many speech processing techniques are plagued for such interferences. Traditional approach to co-channel speech is to attempt to extract the speech of the speaker of interest (target speech) from other (interfering) speech. Usable speech extraction is a novel concept of processing degraded speech data. The idea of usable speech is to identify and to extract portions of degraded speech that are considered useful for various speech processing systems.

Speaker identification system [2-4] needs portions of speech that contain speaker characteristics, which are unique to the individual speakers, classifiable and long enough for the systems to make the decision [1]. These portions of speech are termed as usable speech and defined as consecutive frames of speech that are minimally corrupted by interfering speech. Due to the nature of human voice, a speech utterance contains voiced parts, unvoiced parts and silence; after mixing the two speech signals, there are segments of the co-channel speech that contain only one speaker's voiced part or one speaker's voiced part plus another speaker's unvoiced part, the latter usually having much lower energy. Yantorno [5] performed a study on co-channel speech and concluded that the Target-to-Interferer Ratio (TIR) was a good measure to quantify usability for speaker identification. Usable segment extraction is based on a power ratio of the target speech to the interfering speech. This ratio is expressed as TIR (Target to Interferer Ratio, in dB). The ratio can be expressed for entire utterances or individual frames of speech. For usability, previous experimentation has shown that for frames above 20 dB TIR is considered usable, and that lower 20 dB TIR is considered

unusable segments. The concept of "usable" segments relies on the fact that for any given time frame, the energy of each speaker may be different. The usable speech concept takes advantage of the situation when the energy of the primary speaker is much greater than the energy of the interfering speaker for a given frame. Different criteria are developed to extract usable speech in co-channel speech [6-7]. Criteria such frame-level TIR or spectral autocorrelation ratio. These studies find that voiced segments contain most of the information for SID, and according to these criteria, a significant amount of co-channel speech can be considered usable for SID. Frame TIRs are easily calculated with premixing speech utterances, and usable speech extracted based on a TIR threshold retains frames where target speaker is much stronger in terms of overall energy than the other. Spectral autocorrelation ratio estimates the ratio between dominant peak and valley in autocorrelation of a spectral frame. This ratio is used to determine whether a frame is usable, meaning the spectrum is well structured (single-speaker speech), or not. This approach is simple and effective and shows a substantial improvement in SID performance. Hence, a number of methods for usable speech detection which refer to the TIR have been developed and studied under co-channel condition [8][9]. In these methods, usable speech frames are composed of voiced speech. In [8], the Spectral Autocorrelation Ratio method was developed to detect usable speech segments. This takes advantage of the structure of voiced speech in the frequency domain. In [9], the Peak difference autocorrelation of wavelet transform (PDAWT) method is applied in order to detect pitch information in usable speech. This method applies autocorrelation on approximation component obtained by filtering co-channel speech at one discrete wavelet transform (DWT) scale. These methods show that the speaker identification system achieves approximately 80% of correct identification when the overall TIR is 20 dB.

In our previous work [10-11], we have developed multi resolution dyadic wavelet (MRDWT) method to detect usable speech. MRDWT method is a linear multi-scale decomposition which applies discrete wavelet transform (DWT) iteratively to detect pitch periodicity. We are motivated by detecting pitch information in all lower frequency sub-bands of co-channel speech. In fact, usable frames are characterized by periodicity features. The MRDWT method gives good hits percentage. The detected usable segments are separated in time and need to be organized into speaker streams. Recently, we have proposed in [12] a speaker assignment system that organizes usable speech segments under co-channel conditions. Then, usable segments are assigned to two speaker groups, corresponding to the two speakers in the mixture. Finally, speakers are identified using the assigned segments.

In this paper, we propose to evaluate co-channel speaker identification system. MRDWT is used for co-channel speech processing. Evaluation is performed on TIMIT database. This paper is organized as follows. In section 2, the linear multi-scale method is presented. In section 3, the co-channel speaker identification system is presented. The experimental results of the proposed algorithm in this paper and the target speaker identification accuracy are also presented in section 4. Finally, our work of this paper is summarized in the last section.

2. LINEAR MULT-SCALE DECOMPOSITION FOR USABLE SPEECH DETECTION

Usable frames are characterized by periodicity features. These features should be located in low-frequency band that includes the pitch frequency. Linear Multi-scale decomposition based on DWT is applied iteratively in order to determine the suitable band for periodicity detection. In this band, periodicity features are not much disturbed by interferer speech in case of usable segments. In case of unusable frames, it is not possible to detect periodicity in all lower sub-bands. At each iteration, autocorrelation is applied to the approximation coefficients in order to detect periodicity [11]. Three dominated local maxima are determined from the autocorrelation signal with a peak-picking algorithm which uses a threshold calculated from local maxima amplitudes. A difference of autocorrelation lag between the first and second maximum and between the second and third maximum is determined. If this difference is less than the threshold, periodicity is detected and co-channel speech segment is classified as usable. This threshold is empirically fixed according to the best evaluation results. The optimum threshold value of 8 samples is chosen at 16 kHz sampling frequency. If at this iteration, periodicity is not detected, DWT is applied to approximation signal in order to detect hidden periodicity feature in finer band frequency. For unusable frames, it is not possible to detect periodicity in all lower sub-bands. A maximum of 4 iterations are allowed. This limit is fixed based on pitch band. The lowest band should correspond to pitch band.

Figure 1 corresponds to a usable frame for male-male co-channel speech. In this case, periodicity is detected only at scale 3.

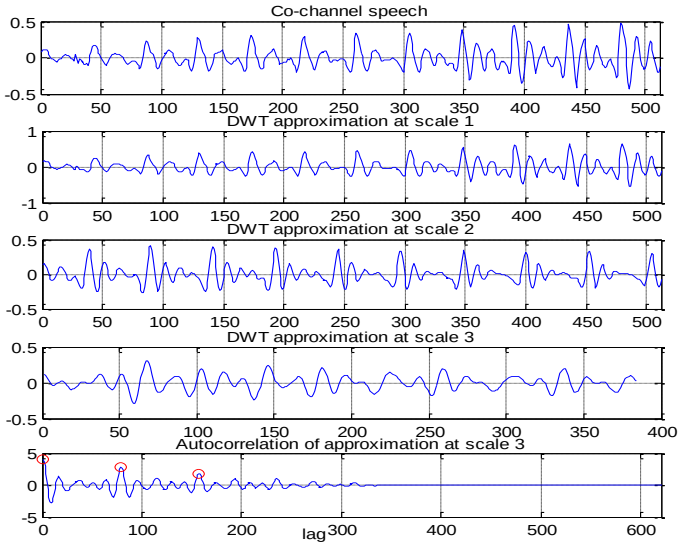

Figure 1. Analysis of a Usable Speech Frame for Male-Male Co-Channel Speech Up to Scale 3

3. CO-CHANNEL SPEAKER IDENTIFICATION SYSTEM

In order to identify the target and the interferer speakers, the detected usable segments are organized into two speaker streams by a speaker assignment system [12]. The speaker assignment system organizes usable speech segments under co-channel conditions. It has extended probabilistic framework of traditional SID to co-channel speech. It uses exhaustive search algorithm to maximize the posterior probability in grouping usable speech. Then, usable segments are assigned to two speaker groups, corresponding to the two speakers in the mixture. The two speaker streams are used as input for a baseline speaker identification system.

3.1. Speaker Assignement

In speaker identification system, discrimination between speakers is based on posterior probability. The goal is to find the speaker model reference in the set of speaker models $\Lambda=\{\lambda 1,\lambda 2,\ldots, \lambda k\}$, that maximizes the posterior probability for an observation sequence $O=\{o1,o2,\ldots,oM\}$ [2]. Cepstral features, such as mel-frequency cepstral coefficients (MFCCs), are used as observations for speech signals. The goal in co-channel attempts to find two speaker models that maximize the posterior probability for the observations.

In [12], we have proposed a speaker assignment system that organizes usable speech segments under co-channel conditions. We have extended probabilistic framework of traditional SID to co-channel speech. For a co-channel mixture, our usable speech extraction method extracts N consecutive speech segments, $X= \{S_1,S_2,\ldots,Si,\ldots,S_N\}$.

Usable segments are organized into two speaker streams because in co-channel speech one speaker can dominate in some portions and be dominated in other portions. For example, a possible segment assignment may look like $S_1^0,S_2^1,\ldots,S_i^0,\ldots,S_N^1$, where superscripts, 0 and 1, do not represent the speaker identities but only indicate that the segments marked with the same label are from the same speaker. In [12], we have demonstrated that probability posterior can be re-written for co-channel speech as:

$$P\left(X^0, X^1 \mid \lambda_I, \lambda_{II}\right) = \prod_{S_i \in X^0} P\left(S_i \mid \lambda_I\right) \prod_{S_j \in X^1} P\left(S_j \mid \lambda_{II}\right)$$

The probability of having a segment S from a pre-trained speaker model λ is the product of likelihoods of that speaker model generating each individual observation x of the segment.

Sirigos et al [13] and Lovekin et al [1] have shown that voiced speech plays a dominant role in speaker recognition. The idea of using only the voiced part of speech signal is based on the fact that voiced speech segment contain the most significant speaker identification as opposed to other speech segment.

When voiced only segments were used for training and testing approximately 80% speaker identity accuracy was achieved. Therefore, we propose to use voiced frame in training. Observations are extracted from voiced frame by MFCCs. Speakers models are formed with 16-mixture GMMs. We employ exhaustive search algorithm to find correspondent speaker. In implementation, the real computation time is longer. It can be further reduced by storing all the likelihood scores of a segment given a model in the memory as a table and looking up a score from the table when needed.

3.2. Baseline SID System

The SID is performed with a baseline system [2-3]. Modeling is assured by Gaussian Mixture Model (GMM) and estimated through the Expectation Maximization (EM) algorithm that maximizes the likelihood criterion. A set of 16 mixtures are used for speaker model. In our experiment, we use the classical parameterization based on 16 Mel Frequency Cepstral Coefficients (MFCC). These coefficients are computed from the speech signal every 10 ms using a time window of 25 ms. Each feature vector is presented by the middle windows of every utterance. Speaker model is trained using the EM algorithm with the features calculated from training samples. In testing phase, the organized usable speech, with speaker assignment system, are used as test speech samples for SID system. The same features are derived from the test speech samples and are input to every speaker's GMM. The speaker with the highest likelihood score represents the identified speaker. Here, speaker identification experiments are close-set and text-independent.

4. EXPERIMENT AND RESULT

Speech data from the TIMIT database was used for all the simulation experiments. The speaker set is composed of 38 speakers from the "DR1" dialect region, 14 of which are female and the rest are male. For each pair of speaker the TIR is calculated as the energy ratio of the target speech over the interference speech. Three different sets of co-channel speech are considered: male-male, female-female, and male-female. Speech signals are scaled to create the mixtures at different TIRs: -20 dB, -10 dB, -5 dB, 0 dB, 5 dB, 10 dB and 20 dB.

4.1. MRDWT Evaluation

The Target to Interferer Ratio TIR measure is used to label voiced frames as usable or unusable. For usability decision, frames that have above 20 dB TIR are considered as usable. Evaluation is based on hits and false alarms percentages.

The performance of proposed method is given in Table 1. We compare the proposed method with related approaches in [9]. On average the MRDWT method detects at least 95.76% of the usable speech with a false alarm rate of 29.65%. Peak difference autocorrelation of wavelet transform (PDAWT) method [9] is based on pitch information detection. This method applies DWT once only to co-channel speech to detect pitch information. On average the PDAWT method detects at least 80% of the usable speech with a false alarm rate of 30%.

Table 1. Results of PDAWT and MRDWT Method for Usable Speech Detection

	PDAWT		MRDWT	
Co-channel	% Hit	% False alarm	% Hit	% False alarm
Female-Female	82.0	32.3	93.02	32.37
Male-Male	80.5	30.6	98.46	28.93
Male-Female	81.3	29.6	95.80	27.66
Average	81.2	30.8	95.76	29.65

MRDWT achieve a maximum of detected usable speech compared to PDAWT. We consider the effectiveness of linear multi-scale decomposition by MRDWT to increase the percent of hit.

4.2. Speaker Identification Evaluation

To demonstrate the usefulness of our proposed method, usable speech is assigned into streams by our speaker assignment system. If the target speaker is of interest, then the speech signal from the other speaker is considered noise. We choose the target speaker SID as our evaluation criterion. Figure 2 shows combination of co-channel female-male. We show the correspondent assignment to speaker 1 and 2

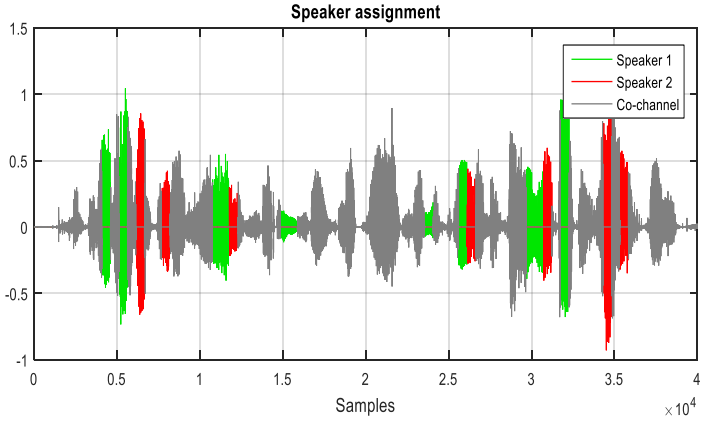

Figure 2. Speaker Assignment for Female-Male Co-Channel

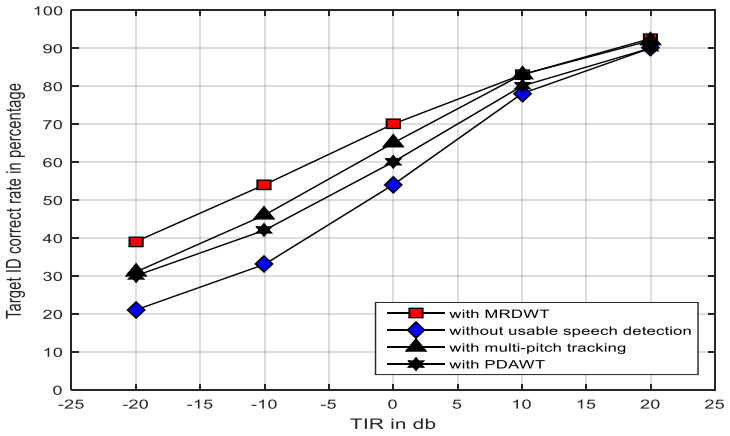

Figure 3. Performance of the Proposed Speaker Identification under Co-Channel Conditions Compared with Related Methods

Figure 3 gives the target SID correct accuracy for our proposed method and related method. In [14] Shao proposed robust pitch tracking method to extract usable speech for speaker identification task. Based on pitch information, this method extracts the usable speech segments that consist of only one speaker's pitch and feed them into a speaker identification system. Target SID correct rate are given before and after usable speech extraction. It's clear from Figure 3 that the MRDWT performs significantly better than Multi-pitch tracking and PDAWT usable speech methods. MRDWT, Multi-pitch tracking and PDAWT usable speech methods and speaker assignment system improves significantly SID performances under co-channel situations. The accuracy degrades sharply when TIR decreases because the target speech is increasingly corrupted. The first observation from the figure is that, under co-channel situations, the target SID correct rate with usable speech detection is better than the target SID correct rate without usable speech detection. The proposed usable speech detection improves speaker identification performance. Secondly, the improvements are consistent across all TIR levels. Performance improvement increases at higher TIRs because the target speaker dominates the mixture. However, target speaker is dominated by interference at lower TIRs, resulting in lower performance after usable speech extraction.

5. CONCLUSION

In this paper, we have employed a new method for usable speech extraction to improve speaker identification under co-channel speech. Usable speech is extracted based on pitch information obtained by linear multi-scale decomposition MRDWT. Our usable speech extraction method produces segments useful

for co-channel SID across various TIR conditions. MRDWT achieves a good percent of usable speech detection. In comparison with PDWAT method, our proposed method achieves a good percent of correct detection. We consider the effectiveness of multi-scale decomposition to extract usable speech. Usable segments are assigned to two speaker groups, corresponding to the two speakers in the mixture. Organized usable speech are used as input to speaker identification system. We have shown that the proposed usable speech detection achieves good SID performance and it performs significantly better than without usable speech detection. SID performance degrades when TIR decreases because the target speech is increasingly corrupted by interferer speech.

REFERENCES

[1] J. Lovekin, R.E. Yantorno, S. Benincasa, S. Wenndt, M. Huggins, *Developing usable speech criteria for speaker identification*, Proc. ICASSP 2001– P. 421-424.

[2] D.A. Reynolds. *Automatic speaker recognition using Guassian mixture speaker model*, Lincoln Lab. J., 1995 – Vol. 8 – P. 173-192.

[3] Kayode Francis Akingbade, Okoko Mkpouto Umanna, Isiaka Ajewale Alimi."Voice-Based Door Access Control System Using the Mel Frequency Cepstrum Coefficients and Gaussian Mixture Model". *International Journal of Electrical and Computer Engineering (IJECE)*, Vol 4 No 5, 2014.

[4] Di Wu, China; Jie Cao; HuaJin Wang, "Speaker Recognition Based on ivector and Improved Local Preserving Projection". *TELKOMNIKA Indonesian Journal of Electrical Engineering*, Vol 12 No 6, 2014 pages 4299-4305.

[5] R.E. Yantorno, "*Co-channel speech study, final report for summer research faculty program*", Tech. Rep., Air Force Office of Scientific Research, Speech Processing Lab, Rome Labs, New York, 1999.

[6] Yantorno, R.E, Method for improving speaker identification by determining usable speech, *J. ACOUST. SOC. AM.*, 2008 – Vol. 124, issue 5.

[7] Brett Y Smolenski and Ravi P. Ramachandran, Usable Speech processing: a filterless approach in the presence of interference; *IEEE Circuits and Systems Magazine*, (2011).

[8] K.R. Krishnamachari, R.E. Yantorno, D.S. Benincasa and S.J. Wenndt, "Spectral autocorrelation ratio as a usability measure of speech segments under cochannel conditions", *IEEE International Symposium Intelligent Sig. Process. and Comm Sys*., (2000).

[9] Kizhanatham, R.E. Yantorno, *Peak Difference Autocorrelation of Wavelet Transform Algorithm Based Usable Speech Measure*, 7th World Multi-conference on Systemic, Cybernetics, and Informatics, 2003.

[10] W.Ghezaiel, A.Ben Slimane, E.Ben Braiek. Usable speech detection for speaker identification system under co-channel condition, JTEA 2010 Tunisia.

[11] Wajdi Ghezaiel, Amel Ben Slimane, Ezzedine Ben Braiek. Evaluation of a multi-resolution dyadic wavelet transform method for usable speech detection, *waset journal*. 2011 – Vol.79 – P. 829-833.

[12] Wajdi Ghezaiel, Amel Ben Slimane, Ezzedine Ben Braiek. "Usable Speech Assignment for Speaker Identification under Co-Channel Situation". *International Journal of Computer Applications*, 59(18): 7-11, December 2012.

[13] J. Sirigos, N. Fakotakis, G. Kokkinakis: "*A comparison of several speech parameters for speaker independent speech recognition and speaker recognition*", in proc. Eurospeech'95, Madrid, Spain, (1995) 18-21.

[14] Shao Y. and Wang D.L. (2003): *Co-channel speaker identification using usable speech extraction based on multi-pitch tracking*. Proceedings of ICASSP-03, vol. II. 205-208.

BIOGRAPHIES OF AUTHORS

Wajdi Ghezaiel born in Tunis (Tunisia), he received the bachelor degree from the high school of sciences and techniques of Tunisia since 2001, the Master Degree and the Ph.D degree in Signal processing from High School of Sciences and Techniques of Tunis respectively in 2004 and 2014. He belongs to the CEREP group in the high school of sciences and techniques of Tunisia. Dr Ghezaiel has published over 11 scholarly research papers in many journal and international conferences and he is already supervising over ten masters and engineer application projects. His research interests are focusing on signal, voice recognition and filtering.

Amel Ben Slimane is currently Assistant Professor of telecommunications at the National School of Computer Sciences (Ecole Nationale des Sciences de l'Informatique, ENSI), University of Manouba, Tunisia since 2003. She received her Engineering degree and Ph.D in electrical engineering both from National School of Engineering of Tunis (Ecole Nationale d'Ingénieurs de Tunis, ENIT) respectively in 1985 and 2003. Her research interests focus on signal processing. Particularly, She works on speech processing. The results of her research work have been published in many international conferences.

Ezzedine Ben Braiek obtained his HDR on 2008 in Electrical Engineering from ENSET Tunisia. He is, presently, professor in the department of electrical engineering at the technical university ESSTT and manager of the research group on signal and image processing at the CEREP. His fields of interest include automatics, electronics, control, computer vision, image processing and its application in handwritten data recognition.