3rd International Conference on Automation, Control, Engineering and Computer Science (ACECS'16)
Proceedings of Engineering & Technology (PET)
pp. 553-557

USABLE SPEECH ASSIGNMENT FOR SPEAKER IDENTIFICATION SYSTEM

Wajdi Ghezaiel[#1], Amel Ben Slimane[*2], Ezzedine Ben Braiek[#3]

[#]*CEREP-ENSIT, University of Tunis*
Tunis, Tunisia
[1]wajdi.ghezaiel@gmail.com
[3]Ezzedine.Benbraiek@esstt.rnu.tn
[*]*ENSI, University of Mannouba*
Mannouba, Tunisia
[2]Amel.benslimane@ensi.rnu.tn

***Abstract*— Usable speech criteria are proposed to extract minimally corrupted speech for speaker identification in co-channel speech. Extracted usable segments are separated in time and need to be organized into speaker streams for speaker identification system. In this paper, we focus to organize extracted usable speech segment into a single stream for the same speaker by speaker assignment system. We extend probabilistic framework for speaker identification system to co-channel speech. We propose to use only voiced part of speech signal in training phase. In fact voiced speech segment contain the most significant information for speaker identification as opposed to other speech segment. The system is evaluated on co-channel speech and results show a significant improvement across various Target to Interferer Ratios for speaker identification.**

Keywords*— co-channel, usable speech; speaker assignment; speaker identification.

I. INTRODUCTION

The co-channel speech is a combination of speech utterances over a single communication channel. The traditional approach to co-channel speech is attempted to extract the speech of the speaker of interest (target speech) from other (interfering) speech. Research has been carried out for decades to extract one of the speakers from co-channel speech by either enhancing target speech or suppressing interfering speech.

Usable speech is a novel approach to the co-channel speech processing problem. The idea of usable speech is to identify and extract portion of degraded speech that are considered useful for various speech processing and concluded that are considered useful for various speech processing system. Yantorno [1][2] 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.

In our previous work, we have proposed multi resolution dyadic wavelet (MRDWT) [3, 4, 5] and empirical mode decomposition (MREMD) [6, 7, 8, 9] methods to detect usable speech. MRDWT method applies dyadic 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 MREMD method for usable detection by empirical mode decomposition, we use EMD to decompose voiced co-channel speech into a linear combination of two components. The first component called intrinsic mode function (IMF) is ranging from the high-frequency band and so-called detail. The second component called residue is ranging to low-frequency band and so-called approximation. Autocorrelation is applied on approximation to detect pitch information. Pitch information is tracked by applying EMD iteratively to achieve the dominant frequency band of pitch. Hence, this approach could easily extract periodicity feature from alls lower frequency sub-bands of co-channel speech.

In co-channel speech, either speaker can randomly appear as the stronger speaker or the weaker one at a time. Hence, the extracted usable segments are separated in time and need to be organized into speaker streams for speaker identification system. Morgan et al. [10] proposed a speaker assignment algorithm using a maximum likelihood criterion to group recovered signals into two speaker streams, one for the target and the other for the interferer. The assignment algorithm groups the individual frames by examining the pitch and spectral continuity for consecutive voiced frames, and comparing the spectral similarity of the onset frame of a voiced segment with recently assigned frames using a divergence measure proposed by Carlson and Clement [11], which is the symmetrized Kullback–Leibler divergence [12].

In this paper, we propose a speaker assignment system that organizes usable speech segments under co-channel conditions. We use extended probabilistic framework of traditional speaker identification system to co-channel speech.

ISSN: 2356-5608

We employ 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. Finally, speakers are identified using the assigned segments.

Evaluation of this method is performed on TIMIT database referring to the TIR measure. Co-channel speech is constructed by mixing all possible gender speakers. Discussion of the merits and limitations of the proposed method are provided basing on evaluation results.

II. 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 $O=\{o_1,o_2,\ldots,o_M\}$ [13][14]. Cepstral features, such as mel-frequency cepstral coefficients (MFCCs), are used as observations for speech signals. The speaker identification decision rule is

$$\lambda_I^{'} = \text{argmax } P(\lambda \mid O)\,;\ \lambda \epsilon\, \Lambda \qquad (1)$$

The goal in co-channel attempts to find two speaker models that maximize the posterior probability for the observations.

In [15], we have proposed a speaker assignment system that organizes usable speech segments under co-channel conditions. We have extended probabilistic framework of traditional speaker identification system to co-channel speech. For a co-channel mixture, our usable speech extraction method extracts N consecutive speech segments, $X= \{S_1, S_2,\ldots, S_i,\ldots, S_N\}$. Posterior probability can be extended as follows:

$$\lambda_I^{'}, \lambda_{II}^{'} = argmax P(\lambda_I, \lambda_{II} \mid X); \lambda_I, \lambda_{II} \in \Lambda \qquad (2)$$

which is to provide a pair of speaker models, $\lambda_I^{'}$ and $\lambda_{II}^{'}$, from the speaker set Λ that maximize the posterior probability given usable speech segments. Usable segments must be 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. Therefore, the objective of speaker assignment is tracking a pair of speaker models, $\lambda_I^{'}$ and $\lambda_{II}^{'}$, together with a segment assignment, $y^{'}$, that maximize the posterior probability:

$$\lambda_I^{'}, \lambda_{II}^{'}, y^{'} = argmax\ P(\lambda_I, \lambda_{II}, y \mid X); \lambda_I, \lambda_{II} \in \Lambda, y \in Y \qquad (3)$$

Y is the assignment space, which includes all possible assignments (labeling) of the segments.

The decomposition of the posterior is analogous to speech recognition based fragment grouping in [16], and model based sequential organization in co-channel speech in [17].

The objective then becomes finding two speakers and an assignment that have the maximum probability of assigned usable speech segments given the corresponding speaker models as follows:

$$\lambda_I^{'}, \lambda_{II}^{'}, y^{'} = argmax \mathrm{P}\left(\mathrm{X} \mid \mathrm{y}, \lambda_I, \lambda_{II}\right) \qquad (4)$$

Given y, the labeling, we denote X^0 as the subset of usable speech segments labeled 0, and X^1 the subset labeled 1. Since X^0 and X^1 are complementary, the probability term can be written as follows:

$$\lambda_I^{'}, \lambda_{II}^{'}, y^{'} = argmax \mathrm{P}\left(\mathrm{X} \mid \mathrm{y}, \lambda_I, \lambda_{II}\right) \qquad (5)$$

The y term is dropped from the above equation because the two subsets already incorporate the labeling information. Assuming that any two segments, S_i and S_j, are independent of each other given the speaker models and that segments with different labels are produced by different speakers, the conditional probability in (5) can be written as

$$\mathrm{P}\left(\mathrm{X}^0, X^1 \mid \lambda_I, \lambda_{II}\right) = \mathrm{P}\left(\mathrm{X}^0 \mid \lambda_I, \lambda_{II}\right) \mathrm{P}\left(X^1 \mid \lambda_I, \lambda_{II}\right)$$

$$\mathrm{P}\left(\mathrm{X}^0, X^1 \mid \lambda_I, \lambda_{II}\right) = \prod_{S_i \in X^0} \mathrm{P}\left(\mathrm{S}_i \mid \lambda_I\right) \prod_{S_j \in X^1} \mathrm{P}\left(\mathrm{S}_j \mid \lambda_{II}\right) \qquad (6)$$

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, assuming the observations are independent of each other. In other words

$$\mathrm{P}(\mathrm{S} \mid \lambda) = \prod_{x \in S} \mathrm{p}(\mathrm{x} \mid \lambda) \qquad (7)$$

The goal is to find two speakers and one assignment that yield the maximal probability using (7). Given the extracted usable speech segments and individual speaker's models, the problem is to find two speakers in space Λ and Y that maximize probability in (7). The brute-force way to find the maximum is exhaustive search.

Sirigos et al [18] and Lovekin et al [2] 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.

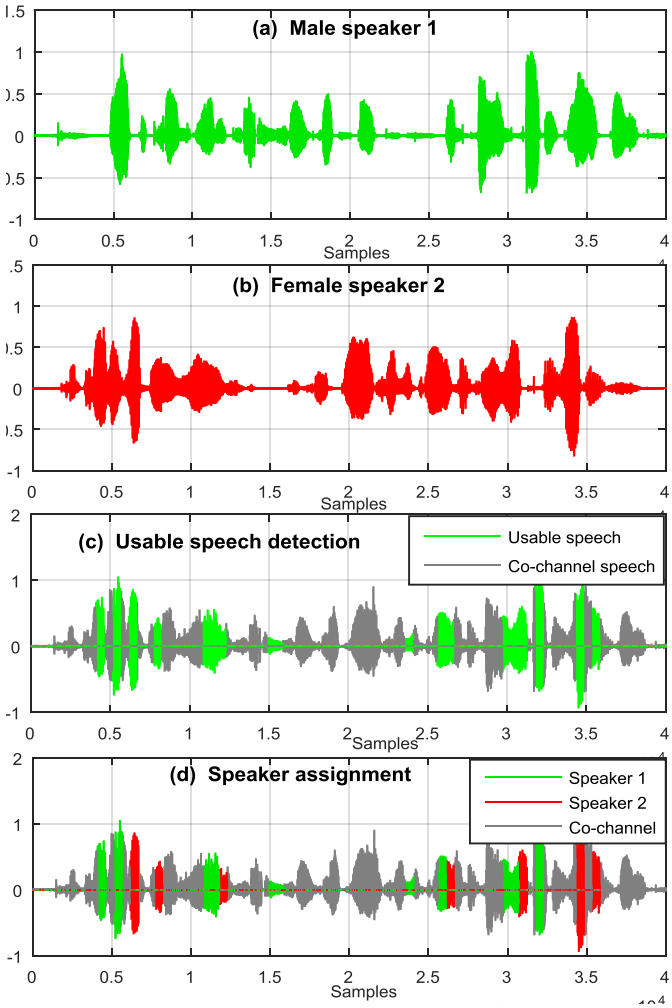

Fig. 2. Usable speech detection and speaker assignment for Male-Female co-channel

III. 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. Each speaker has 10 utterance files, ranging from about 1.5 sec to 6.2 sec in length. For each speaker, 5 out of 10 files are used for training and the remaining 5 files are used to create co-channel mixtures for testing. For each speaker deemed as the target speaker, 1 out of 5 test files is randomly selected and mixed with randomly selected files of every other speaker, which are regarded as interfering utterances. For each pair the TIR is calculated as the energy ratio of the target speech over the interference speech. 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. For example, 0 dB TIR means that the overall energy of target is equal to that of interference. Three different sets of co-channel speech are considered: male-male, female-female, and male-female.

TABLE 1 Target speaker identification accuracy

TIR Level (dB)	-20	-10	0	10	20
Target Speaker Identification (%)	39	54.7	70	83.5	92

A. Speaker assignment evaluation

The developed speaker assignment system is performed based on (7). We use 16 MFCCs coefficients as speaker features. We employ only voiced frame for training. For each speaker, we employ 5 out of 10 files in training phase. Models are formed with 16-mixture GMMs, which are trained using the EM algorithm [14] from the training samples.
Figure 2 shows usable speech detection and the correspondent assignment. We note that given assigned speech segment correspond to original speech of each speaker.

Here, we evaluate the performance of our speaker assignment system. For this evaluation, we only consider co-channel mixtures with overall TIR equal to 0 dB to simulate real co-channel situations. Our speaker assignment system achieves 86.4% correct assignment rate. It reflects the effectiveness of using voiced segment as speaker characteristics for speaker assignment. Performance is compared to sequential organization method proposed par Shao [17]. This method achieves 77.4%. Comparison with sequential organization improves the effectiveness of voiced segment to model speaker characteristics for speaker assignment.

B. Speaker identification evaluation

The speaker identification is performed with a baseline system [14]. Modeling is assured by Gaussian Mixture Model (GMM) and estimated through the Expectation Maximization (EM) algorithm that maximizes the likelihood criterion. A 16 mixture is 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 window. 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 is used as test speech samples for speaker identification 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. We choose the target speaker identification as our evaluation criterion.

Target speaker identification accuracy for different TIR level is given in Table 1. Usable speech extraction and our speaker assignment system improve significantly speaker identification performances. Performance improvement increases at higher TIRs because the target speaker dominates the mixture. The accuracy degrades sharply when TIR decreases because the target speech is increasingly corrupted.

IV. Conclusion

In this paper, we have proposed a speaker assignment system to organize extracted usable speech. Usable segments are assigned to two speaker groups, corresponding to the two speakers in the mixture. We have extended the probabilistic framework of traditional speaker identification system to co-channel speech. We have employed exhaustive search algorithm to maximize the posterior probability in grouping usable speech. Only voiced frame are used for training. The proposed speaker assignment achieves good results in organizing usable speech to corresponding speakers. Comparison with sequential organization improves the effectiveness of voiced segment to model speaker characteristics for speaker assignment. The developed speaker assignment system can be extended to touch on multi-talker condition. We can replace the speaker pair with a speaker triplet, a speaker quadruplet.

Organized usable speech was used as input to speaker identification system. Combination of usable speech and speaker assignment improve the speaker identification performance in co-channel conditions. We have shown that the proposed speaker assignment achieves good speaker identification system performance.

References

[1] Yantorno, R. E, "Method for improving speaker identification by determining usable speech" Journal of the Acoustical Society of America. 124 (2008).

[2] J. Lovekin, R. E. Yantorno, S. Benincasa, S. Wenndt and M.Huggins, "Developing usable speech criteria for speaker identification," Proc. ICASSP (2001) pp. 421-424.

[3] W.Ghezaiel, A.Ben Slimane, and E.Ben Braiek, "Usable speech detection for speaker identification system under co-channel conditions", International conference on electrical system and automatic control JTEA 2010 Tunisia.

[4] Wajdi GHEZAIEL, Amel BEN SLIMANE, Ezzedine BEN BRAIEK, «Evaluation of a multi-resolution dyadic wavelet transform method for usable speech detection», World Academy of Science, Engineering and Technology Journal WASET, pISSN 2010-376X, eISSN 2010-3778, (2011) 829-833.

[5] Wajdi GHEZAIEL, Amel BEN SLIMANE, Ezzedine BEN BRAIEK, « Linear Multi-Scale Decomposition for Co-Channel Speaker Identification System» International conference on Automation, Control, Engineering and Computer Science (ACECS 2015), Sousse, Tunisia, March 2015. Proceeding of Engineering and Technology (PET).

[6] Wajdi GHEZAIEL, Amel BEN SLIMANE, Ezzedine BEN BRAIEK (2013) Usable speech detection based on empirical mode decomposition. IET Electronic Letters 49 Issue 7: 503-504.

[7] Wajdi GHEZAIEL, Amel BEN SLIMANE, Ezzedine BEN BRAIEK Multi-Resolution Analysis by Empirical Mode Decomposition for Usable Speech Detection. International Multi-Conference on Systems, Signals & Devices, Conference on Communication & Signal Processing, SSD 2013 Tunisia.

[8] Wajdi GHEZAIEL, Amel BEN SLIMANE, Ezzedine BEN BRAIEK (2013) Improved EMD usable speech detection for co-channel speaker identification. Lecture Notes in Computer Science, Advances in Non-Linear Speech Processing, vol. 7911, pp. 184-191.

[9] Wajdi GHEZAIEL, Amel BEN SLIMANE, Ezzedine BEN BRAIEK "Linear vs Nonlinear Multi-Scale Decomposition for Co-Channel Speaker Identification System" International conference: Non Linear Speech Processing 2015, (NOLISP 2015), Vietri Sul Mare, Italy, 18-20 May 2015.

[10] Morgan, D. P., George, E. B., Lee, L. T., and Kay, S. M., Cochannel speaker separation by harmonic enhancement and suppression, IEEE Transactions on Speech and Audio Processing 5 (1997), 407-424.

[11] Carlson, B. A., and Clements, M. A. A computationally compact divergence measure for speech processing, IEEE Transactions on Pattern Analysis Machine Intelligence 13 (1991) 1-6.

[12] Kullback, S. (1968). Information theory and statistics. New York: Dover Publications.

[13] Furui, S., 2001. Digital Speech Processing, Synthesis, and Recognition. Marcel Dekker, New York.J.

[14] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Commun., 17 (1995) 91–108.

[15] W. Ghezaiel, A. Ben Slimane and E. Ben Braiek. "Usable Speech Assignment for Speaker Identification under Co-Channel Situation". International Journal of Computer Applications 59(18):7-11, December 2012.

[16] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sources,"Speech Comm., 2004.

[17] Shao, Y., and Wang, D. L. Model-based sequential organization in co-channel speech. IEEE Transactions on Audio, Speech, and Language Processing, 14 (2006), 289-298.

[18] 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.