https://doi.org/10.33261/jaaru.2024.31.3.005 *Association of Arab Universities Journal of Engineering Sciences (2024)31(3):37–45*

كلية الهندسة - جامعة بغداد

Association of Arab Universities
Journal of Engineering Sciences
مجلة اتحاد الجامعات العربية للدراسات والبحوث الهندسية

جمعية كليات الهندسة
اعضاء اتحاد الجامعات العربية

Supervised Machine Learning for Speaker Diarization by PNCC with LPCC Audio Coefficients

Hasan M. Kadhim [1,*], ***Alaa H. Ahmed*** [2], ***Alaa K. Hassan*** [3], and ***Saad T. Y. Alfalahi*** [4]

[1] *Department of Electrical Engineering, College of Engineering, Mustansiriyah University, Baghdad, Iraq, hasanalmgotir@uomustansiriyah.edu.iq*

[2] *Department of Electrical Engineering, College of Engineering, Mustansiriyah University, Baghdad, Iraq, alaa75hs@uomustansiriyah.edu.iq*

[3] *Department of Electrical Engineering, College of Engineering, Mustansiriyah University, Baghdad, Iraq, alaak_eng@uomustansiriyah.edu.iq*

[4] *Department of Computer Engineering, Madenat Alelem University College, Baghdad, Iraq, saad.t.yasin@mauc.edu.iq*

** Corresponding author: Hasan M. Kadhim, and email: hasanalmgotir@uomustansiriyah.edu.iq*

Published online: 30 September 2024

Abstract— Speaker Diarization is a speech digital signal processing technique that segregates one input observation of *n* multi-speaker signal into an individual speech of those *n* persons. Each segregated signal belongs to one of them plus a bit of error, which is speech that belongs to other speakers. The format of that speech is a dialog because they speak non-simultaneously. By the use of speaker diarization in this research, audio features are extracted from speech. The extraction is the training stage of machine learning. The second classification stage can then decide how to divide these features into those *n* groups. Linear Prediction Cepstral Coefficients (LPCC) and Power-Normalized Cepstral Coefficients (PNCC) are used independently to generate their features. In this paper, the researchers re-combined these LPCC and PNCC features to form a new mixture of features. Improved Euclidian distance facilitates the job of measuring distances to identify who is the nearest label. Because PNCC is a non-inversible transformation, a small frame at the center of a large windowed frame has been regarded (because it has a reasonable weight) to obtain original speech signals. The procedure was efficient for clustering a mixture of two speaker signals, female and male from the TIMIT standard audio library, i.e., successfully recovered each person's individual speech. The average Diarization Error Rate (SDR) objective tests of the recovered speech were 1.8% for the females, 2.9% for the males, and 2.5% for the overall females and males. Compared with other standard research, the improvements were 6.5% for the females, 10% for the males, and 8.8% for all females and males.

Keywords— Speaker Diarization, LPCC, PNCC, Clustering, Diarization Error Rate.

1. Introduction

Suppose there is the following spontaneous speech chat between Girl (G = White color rectangles in Figure 1) and Boy (B = Black color rectangles in Figure 1); first row in Figure 1: At first the Boy (B, the girl is silent) is speaking alone, then the Girl is speaking alone (G, the boy is silent), then both of them the girl with the boy are not speaking (there is a silent period (S) which is without rectangle in Figure 1), then both of them the girl with the boy are speaking simultaneously (Gray color (Gr) in the following figures), then there is a silent period, then the Girl is speaking (G), then both of them the girl with the boy are speaking somnolently (Gr overlapped speech between them), then the Boy (B) is speaking alone, and then both of them are speaking somnolently (Gray). The sequence of the speech/ speaker is: B, G, S, Gr, S, G, Gr, B, and then Gr. When both of the two speakers are talking together at the same time simultaneously, this format of speech is called a Mixture (it is called "cocktail-party problem" in DSP). When one of the two speakers is talking while the other is not, this format of speech is called a Dialog. For the dialog speech format, sometimes there are few-time durations of overlapped-speech between the speakers, which is a mixture speech format. For such multi-speaker/ speech signal processing, the reader can note that there are three following main cases should be resolved [12,15]:

- Overlapped-Speech Detection Process: Its input observation speech signal is a conversation that consists of mixture, dialogue, and overlapped-speech formats speech signals, first row/ Figure 1.

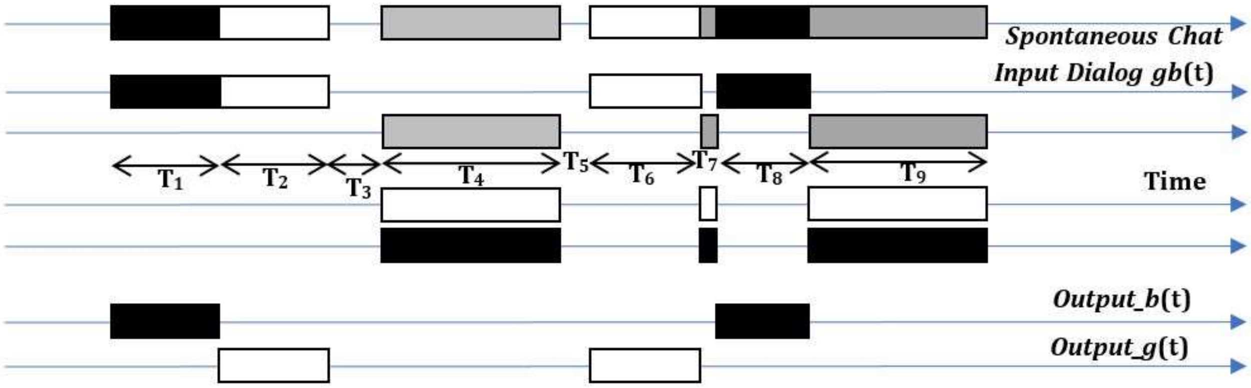

Figure 1: Two speakers (White for Girl and Black for Boy) overlapped-speech detection, speech separation and speaker diarization of the input first row spontaneous speech chat. The waveform is described in the main text.

The job of the detection process is the splitting of that input main signal into two sub-signals: The first sub-signal is a dialog format speech i.e., Black (boy) and White (girl) color which is the second row of Figure 1; and the second sub-signal is a mixture speech format speech, i.e., Gray (the girl and boy simultaneously are speaking) color which is the third row of Figure 1. Almost, Overlapped-speech detection is a supervised Machine Learning (ML) process [3, 4, 7, 20].

- Speech Separation Process: Its input observation speech signal is the mixture format, which is the third row/ Figure 1. The speech separation process tries to retrieve the speech of each speaker alone with a little bit of the other speakers' speech. The error signal should be avoided by the reducing or deleting as much as possible from the desired speech signal of that specific speaker. Those output retrieved speech signals are the fourth and the fifth rows/ Figure 1 (White for girl and Black boy respectively). When the input observation signal does not have any other information or database belonging to the speech/ speakers, it's an unsupervised Machine Learning process. When the input has any other information or database belonging to the speech/speakers, it's a supervised Machine Learning (ML) process. When the process could produce/generate some information or database belongs to the speech/speakers, it's a semi-supervised Machine Learning (ML) process [11, 23, 27].
- Speaker Diarization (SD) Process: Its input observation speech signal is the dialog format (Gray color), which is the second row/ Figure 1. The speaker diarization process tries to retrieve the speech of each speaker alone with a little bit of the other speakers' speech. Those speech are error signals which should be avoided by the reducing or deleting as large as possible from the wanted speech signal of that specific speaker. The output retrieved speech signals are the sixth and the seventh rows/ Figure 1 (Black for boy and White for girl respectively). Speaker diarization is a supervised and an unsupervised ML process [2, 5, 14, 21].

For the literature that were focusing on SD, [11, 12] employed the diarization output of individual speech as a tiny dataset for informed supervised speech separation. Their system provided Source-to-Interference Ratio (SIR), Source-to-Distortion Ratio (SDR) and Source-to-Artifact Ratio (SAR): 9.8, 3.05 and 4.65dB using 0.4% average SD-EMISS algorithms; and 11.1, 1.7, and 2.8dB. They utilized the NNMF algorithms on each sub-band of the filter-bank. In [13], the researchers optimized the clustering stage of SD by k-means several times. The optimization exploited the stochastically oriented properties of the labels. Hybrid Bottom-up with Top-down scenarios used for final stage of SD. The individual recovered waveforms approximated to the original signal with a little bit of errors. In [16], the researchers tackled the SD overlapped-speech issue. For the detection, they designed an architectural Neural Long system with a memory-based (Short-Term). Switch time from speaker to another used in the next stage. They tested their system on standard libraries with DER of 20%. In [10], a large number of SD speakers were chatting in the conversations. The researchers suggested decoding with an encoding approach. DERs of double-speakers were from 2.69 to 9.54 % on different standard datasets. DERs of multi-speakers were from 15.29 to 19.43 %. Authors of [28] suggested a supervised ML SD system. They shared Neural Network for the time interleaved speaker. The ML system used to share the speech clusters between the labels. On the standard speech NIST library, the DER was 7.6%. The clustering process is done on the spectrum.

2. Speaker Diarization

In this article, the researchers focused on the third case problem where one of two speakers is talking while the second is silent during a specific duration, then the second speaker is speaking while the first one is silent. As we mentioned above, the process is called Speaker Diarization Digital Signal Processing. Suppose the speakers are *g* (girl, White color) and *b* (boy, Black color). The input observation signal is gb(t), where:

$$gb(t) = g(t) + b(t) \tag{1}$$

Where: g(t) is available arbitrary in the time domain during the periods, i.e., T_2, T_6, etc., b(t) is available in the time domain during the periods, i.e., T_1, T_8, etc. The g and b are not available during the silence periods T_3, T_5, …etc. Typically, any odd period must be adjacency with an even or silence period; and any even period must be adjacency with an odd or silence period as well. The first output is:

$$Output\_g(t) = g(t) + e_g(t) \qquad (2)$$

And the second output is:

$$Output\_b(t) = b(t) + e_b(t) \qquad (3)$$

Where e_g(t) is undesired error signal which belongs to b(t); and e_b(t) is undesired error signal which belongs to g(t). In the second row/Figure 1, gb(t) is the total input observation waveform. The White waveforms represent g(t) in the seventh row/Figure 1, and the Black waveforms represent b(t). Algorithms of speaker diarization consist of two interconnected sequential stages. The first stage is the speaker segmentation of the conversation between the two or more speakers. The second stage is speaker clustering which identifies the speakers of these segments. The speaker segmentation partitions the chat input speech observation signal into groups of segments. Each speech signal segment group has similar characteristics, specifications and features among these segments' signals. That segmentation process can be done by determining the switching instances from any speech segment to the adjacent speech of another speaker, or to the adjacent silence period. Speaker clustering is a DSP process which identifies or personalizes each speech segment to a specific unique speaker. In machine learning, the label process could indicate to those segments and clusters. During the previous decades, applications of speaker diarization were increasing rapidly. The conversation between the banks narrators with clients is one of them. In Multimedia and TV broadcasting, speaker diarization has significant effects to split the speech of the pundit/ correspondence and the quests during the regular live and recorded meeting and interviews. A lot of practical useful applications exploit speaker diarization [2, 6, 21].

3. Proposed Algorithm

Speaker Segmentation is the first stage of the speaker diarization process for the of the input speech chat for the *g* and *b* speakers. The segmentation process can determine the switching instances and boundaries among sentences, phonemes, syllables, words, and letters of the spontaneous speech chats. The segmentation could be implemented artificially by smart machines and naturally by the human brain. Speaker/speech recognitions and speech perceptions are superset of speech/ speaker segmentations. There are subsequent interconnections between these three areas of audio and speech DSP. For language processing issues, other parameters should be considered such as the Grammer, context, semantics and the likelihoods for statistics and stochastics [13]. For the phenomenon, co-articulation could occur for adjacent sentences and words. The overlapping between these components is a true challenge for researchers in this field. In the previous first and second sections, the researchers described the effects of small-periods overlapping between one or more speakers and how the detection process can resolve this issue. For audio and speech DSP, speaker segmentation exploits the extracted features of the audio signal (in this article, its speech signal). The well-known reliable algorithms to extract the speaker coefficients (features), such as the Linear Prediction Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC), Linear Power-Normalized Cepstral Coefficient (PNCC), and Perceptual Linear Prediction Coefficients (PLPC) [1, 26]. Feature extraction should express specific content/ information from the speech/audio signals. The extraction algorithms can grant specific subsequent models for a speaker and classify the speech/speaker. These extracted coefficients could perform this relevant content/ information by optimize minimizing the intra-speaker of the variability, and the inter-speaker (maximizing) for the variability, ideally. Ordinary speaker diarization process utilize one of these methods to perform the process of speaker segmentation. The main issue with using specific one these methods is the fluctuated performance from speaker to other speaker, from sentence to other sentences, from language to other languages, and form recorded speech at specific conditions to recorded speech at other conditions. To reduce the effects of these fluctuation phenomena, researchers of this article proposed a combination procedure to use two feature extraction methods instead of one. Instead of one coefficient vector per speech frame, the researchers proposed an enlarged-vector which contains features of these two algorithms per that speech frame. The researchers experimented with these reliable algorithms and then chose the following best two combined algorithms:

3.1 *Power-Normalized Cepstral Coefficients (PNCC)*

It is the newest algorithm for speech coefficient extraction developed by Kim and Stern [17, 18]. According to the PNCC algorithm, the speech signal passes through a pre-emphases filter. Then partition the output signal to main frames of 16 to 32 m second duration. Each two adjacent main frames are overlapped by 50% to 75%, i.e., the hopping ratio is 50% to 25% of the main frame period. Each main frame signal values are scaled by the standard Hamming or Hanning window one-by-one. Using Short-Time Fourier Transform (STFT), Time-Frequency domain spectrogram is arranged of all the frames. All positive values of the spectrogram matrix are scaled again by the Gammatone filter-bank. After the previous scaling, the matrix is power-normalized for its average values. The power of each frame value is reduced to 1/15 of its value. Using Discrete Cosine Transform (DCT) frame-by-frame, and then normalize the output average values, Figure 2.

3.2 *Linear Prediction Cepstral Coefficients (LPCC)*

It is the modified version of the standard Linear Predictive Coding (LPC). According to the LPCC algorithm, the speech signal passes through a pre-emphases filter. Then partition the output signal to main frames. Each two adjacent main frames overlapped. The hopping period is the standard DSP duration for speech which is 8 to 16 m second. Each main frame signal values are scaled by the standard Hamming or Hanning window one-by-one. Each main frame is autocorrelated with himself. The output of the autocorrelation is analyzed by the standard LPC. The output of the analyzer is converted by LPCC conversion process, which are the Linear Prediction Cepstral Coefficients values, see Figure 3 [9]. The above configuration (Combination) of the two algorithms to produce double enhanced speech features is efficient for the training phase of the Machine Learning process [6].

3.3 *Speaker Clustering*

The most efficient algorithms for speaker clustering are the two scenarios: Bottom-Up and Top-Down Scenarios. Bottom-up is a more successful scenario for speech and speaker clustering. Iterative calculations are used for both the top-down and bottom-up scenarios. Figure 4 illustrates the starting initial points for each scenario, and the end finalizing points for them. The figure shows the sequence procedure from the starting initial points to the end finalizing points. Through that procedure, specific algorithms should be adopted to support the procedure efficiently. Several algorithms are proposed to implement speech and speaker clustering. For this research, distance measuring was suitable to choose the best label of the observation signal features. Euclidian-distance is used efficiently to measure distances from the features to other features of the combined feature vectors. The previous segmentation facilitates the job of clustering. The calculation to comparator through each 25 to 35 frames' time. That means, the duration of the sequence is 0.8 to 1 second of the input observation signals. Inside the 0.8 to 1 second, Euclidian-distance is helpful to choose the nearest label (small specific frame) that is added to or deleted from the adjacent label (small specific frames). Depending on which scenario has been used, the adding to or deleting from, are executed. Each scenario could be used independently from the other scenario, then the results of them should be regarded using the average values of the calculation rests. The two scenarios could be invoked simultaneously to implement clustering [14, 25].

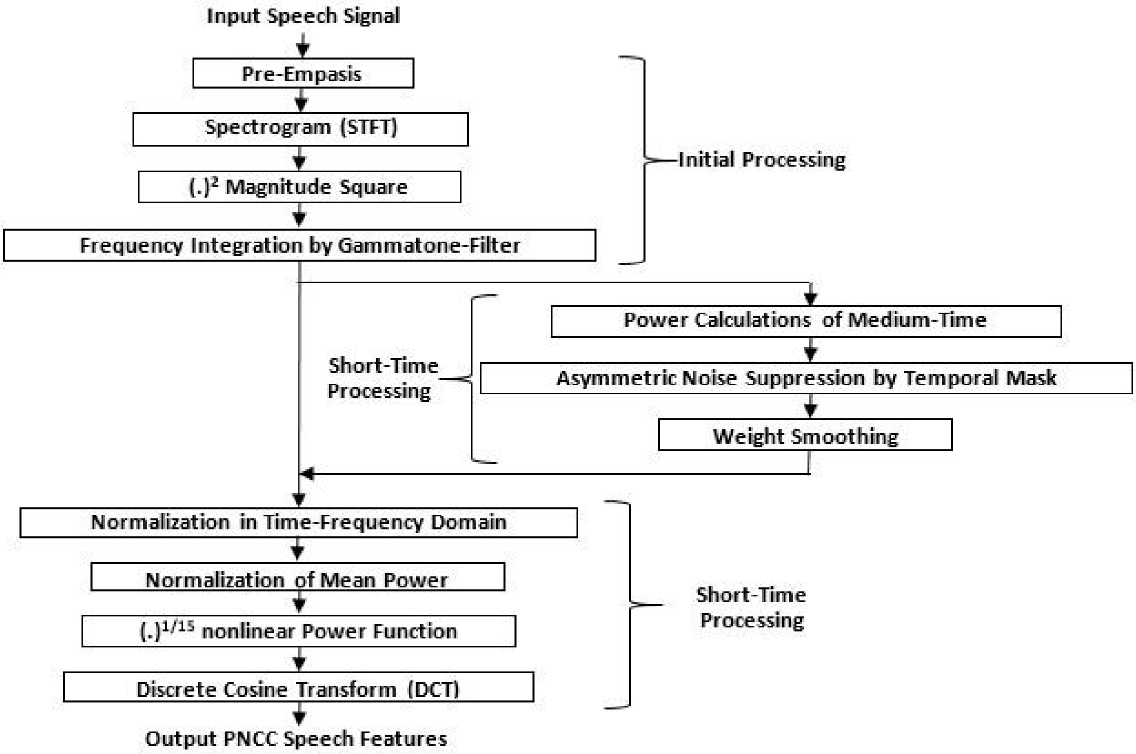
Input Speech Signal
Pre-Empasis
Spectrogram (STFT)
(.)² Magnitude Square
Frequency Integration by Gammatone-Filter
Initial Processing
Power Calculations of Medium-Time
Short-Time Processing
Asymmetric Noise Suppression by Temporal Mask
Weight Smoothing
Normalization in Time-Frequency Domain
Normalization of Mean Power
(.)^1/15 nonlinear Power Function
Discrete Cosine Transform (DCT)
Short-Time Processing
Output PNCC Speech Features

Figure 2: Functional block diagram of the Power-Normalized Cepstral Coefficients (PNCC).

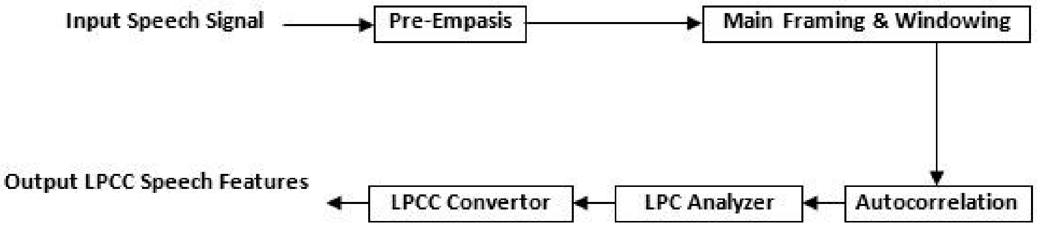

Figure 3: Functional block diagram of the Linear Prediction Cepstral Coefficients (LPCC).

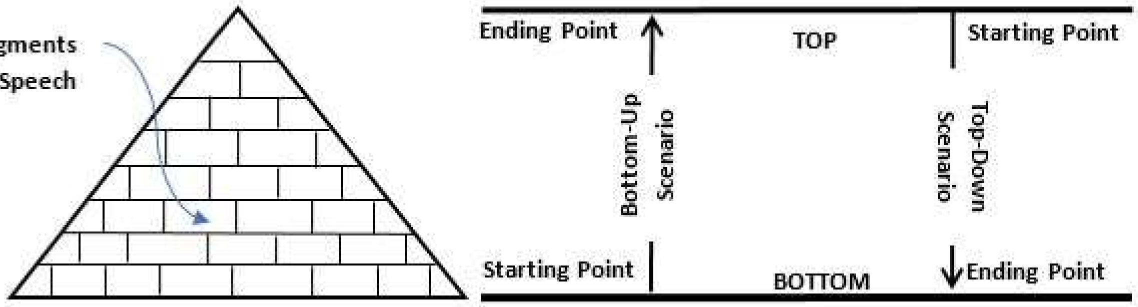

Figure 4: Typical sketch for the two efficient scenarios for the speech/ speaker clustering: the Bottom-Up scenario on the left side, and the Top-Down scenario on the right side.

4. Analysis, Tests and Results

The research is simulated using two speakers, female and male of the standard TIMIT speech library. The time of continuous speaking is about 15 minutes each one. In order to cover possible ranges, data-base prepared with 8, 11.025, and 16khz sampling rates (i.e., 4, 5.5125 and 8khz bandwidth). Energy or power calculated frame by frame to ensure that normalization exists. Neither long, nor short duration of each frame can normalize energy of speech signals, because very long frames (more than several seconds) choice corrects normalization slightly, and very small frames (less than large frame duration) will amplify silence periods. Normalization calculations can be done in the time and/or frequency domain. Dynamic Time Warping (DTW) has been executed using available prepared MATLAB code. More details can obvious knowhow of processing. All speech files have a 16-bit resolution and a 16khz sampling rate. Observation consists of 6 segments belonging to the TIMIT female and male. Pitch Detection Algorithm (PDA) [22] was used to remove silence and unvoiced speech frame-by-frame. 512 point of Hamming window scales large frames of speech. Increment (hopping time) of large frames is 10 msec (the standard processing time for speech). Dividing each speech file into large frames. Extracting features (LPCC and PNCC) of these files, then combining the coefficients of each long frame to arrange independent features matrices. Calculate normalized statistical patterns of a database. Measuring distances of observation to patterns inside long-time (about 5 second) frames. Decision of nearest has effect on decisions of adjacent frames. Because LPCC and PNCC are non-reversible transformations, retrieving only central small frames inside large frames. Assemble the resulting frames to create output speech files (number of output files equals number of speakers). Subjectively and objectively check the output speech signal and the third waveform (lower) is the second output speech signal [8].

The waveforms of all signals, those displayed by the media player (Figure 5), Subjective tests indicate that Diarization of 2 speakers (female and male with 2 segments per each speaker) is very good and output speech is deterministic. In order to test the resulting outputs objectively, Diarization Error Rate (DER) calculates numerical values of such processing errors, where DER is defined as [24]:

$$DER = (FA + MISS + ERROR)/TOTAL \quad (4)$$

Where FA: Total time for speaker (hypothesis not a reference speaker attributed), ERROR: The Total time for the reference speaker (wrong speaker attributed), MISS: The total time for reference speaker time (a hypothesis speaker for not attributed), and TOTAL: The total speech time, which is the sum of the time of the all segments (for the reference speaker), Figure 5, Table 1, Figure 6, Table 2, and Figure 7**.**

Compared with other standard researches' systems, the average Diarization Errors Rates (objective tests) of the recovered speech were 1.8% for the females, 2.9% for the males, and 2.5% for the overall females and males. The improvements in the system were 6.5% for the females, 10% for the males, and 8.8% for all females and males.

In contrast, negative points against this research system are the delays in processing. The system needs about 0.75 second to make a proper derision (Due to the system is a Machine Learning system, it consumes more time to complete the required processing). Another week point is the system is supervised Machine Learning.

When this system can be applied using semi-supervised or unsupervised ML with less time of delay, it will be a very good and reliable system [19, 24, 25].

5. Conclusions

According to subjective tests of different genders and cultures of the listeners, the diarization system recovered the original individual speech of each speaker perfectly with negligible and disregarded errors. The system has a very low ratio of error. Since the system is a Machine Learning, it consumes more time to complete the required processing. The system needs additive to make a proper derision, because it is a supervised ML system. According to the above DER objective tests, the tables and the bars of the tables, the proposed combination of speech features can perform the diarization process efficiently.

Acknowledgements

We thank the Electrical Engineering Department/ College of Engineering/ Mustansiriyah University and Department of Computer Engineering, Madenat Alelem University College, Baghdad, Iraq for the support.

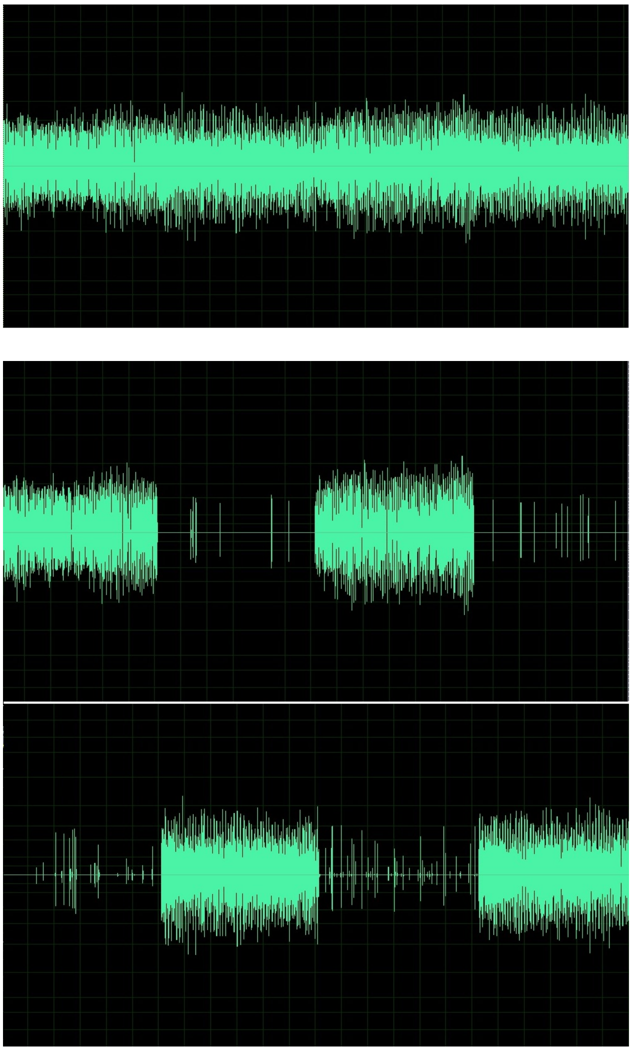

Figure 5: Waveforms: 1st row is the input speech signal for dialog conversation between girl and boy (TIMIT library). 2nd row is the output speech of the girl with error. The 3rd row is the output speech of the boy with error.

Table 1: Average DERs for the LPCC, PNCC and the combined LPCC-PNCC for Females, Males and All.

	LPCC	PNCC	LPCC with PNCC
Females	5.2%	12.5%	2.9%
Males	15.6%	1.3%	1.8%
All	9.3%	7.1%	2.5%

Table 2: Minimum (Min.), maximum (Max.) and average (Av.) DERs% for Females, Males and All.

	Min. DER%	Max. DER%	Av. DER%
Females	0.3%	3.1%	2.9%
Males	0.4%	2.8%	1.8%
All	0.3%	3.1%	2.5%

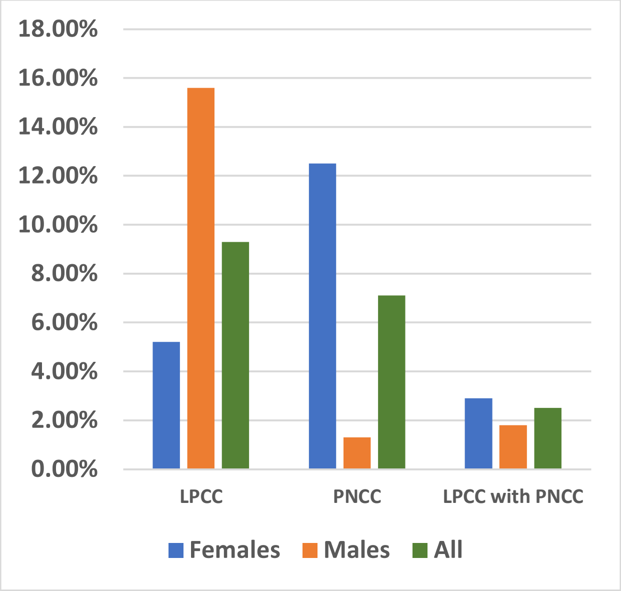

Figure 6: Average %DERs for the LPCC, PNCC and combined LPCC-PNCC for Females, Males and All.

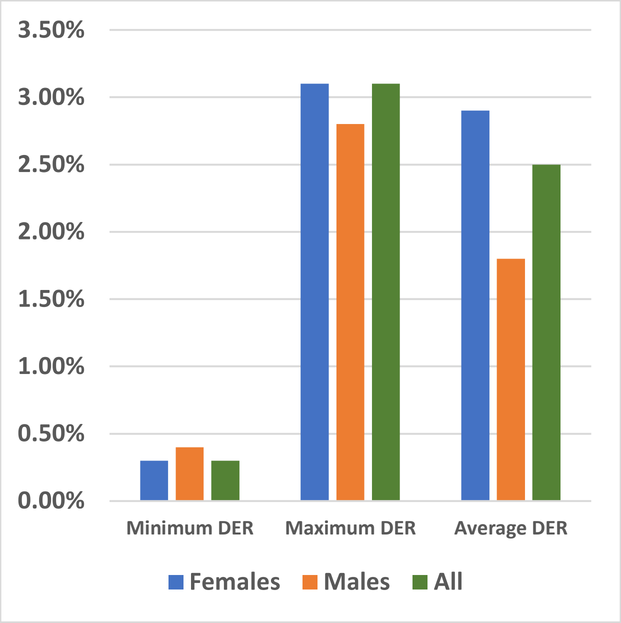

Figure 7: Minimum, maximum and average %DERs for Males (Orange), Females (Blue), and All (Green).

References

[1] Alías, Francesc, Joan Claudi Socoró, and Xavier Sevillano. "A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds." *Applied Sciences* 6, no. 5, 2016 May 12: 143, DOI.

[2] Anguera, Xavier, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. "Speaker diarization: A review of recent research." *IEEE Transactions on audio, speech, and language processing* 20, no. 2, 2012 Jan 23: 356-370, DOI.

[3] Boakye, Kofi, Beatriz Trueba-Hornero, Oriol Vinyals, and Gerald Friedland. "Overlapped speech detection for improved speaker diarization in multiparty meetings." In *2008 IEEE international conference on acoustics, speech and signal processing*, pp. 4353-4356. IEEE, 2008 Mar 31, DOI.

[4] Bullock, Latané, Hervé Bredin, and Leibny Paola Garcia-Perera. "Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7114-7118. IEEE, 2020 May 4, DOI.

[5] Dunbar, Ewan, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. "The zero resource speech challenge 2017." In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 323-330. IEEE, 2017 Dec 16, DOI.

[6] Fujita, Yusuke, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe. "End-to-end neural speaker diarization with permutation-free objectives." *arXiv preprint arXiv:1909.05952*, 2019 Sep 12, DOI.

[7] Ghadanfari, Saleh. "Hierarchical timing in varieties of Kuwaiti Arabic." *PhD diss., Newcastle university*, 2022, URL.

[8] Hernández, Edward L. Campbell, Gabriel Hernández Sierra, and José R. Calvo de Lara. "CENATAV Voice-Group Systems for Albayzin 2018 Speaker Diarization Evaluation Campaign." *Proc. IberSPEECH* 2018, 227-230, DOI.

[9] Hönig, Florian, Georg Stemmer, Christian Hacker, and Fabio Brugnara. "Revising Perceptual Linear Prediction (PLP)." In *Interspeech*, pp. 2997-3000. 2005 Sep, DOI.

[10] Horiguchi, Shota, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Kenji Nagamatsu. "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors." *arXiv preprint arXiv:2005.09921*, 2020 May 20, DOI.

[11] Ismael, Ruaa N., Hasan M. Kadhim, and Susanto B. Sulistyo. "Single-Channel informed spontaneous speech separation by NNMF on a uniform filterbank." In *2023 Al-Sadiq International Conference on Communication and Information Technology (AICCIT)*, pp. 318-323. IEEE, 2023 Jul 4, DOI.

[12] Ismael, Ruaa N., and Hasan M. Kadhim. "NNMF with Speaker Clustering in a Uniform Filter-Bank for Blind Speech Separation." *Iraqi Journal for Electrical & Electronic Engineering* 20, no. 1, 2024 Jun 1, DOI.

[13] Kadhim, Hasan Almgotir, Lok Woo, and Satnam Dlay. "Novel algorithm for speech segregation by optimized k-means of statistical properties of clustered features." In *2015 IEEE International Conference on Progress in Informatics and Computing (PIC)*, pp. 286-291. IEEE, 2015 Dec 18, DOI.

[14] Kadhim, Hasan Almgotir, Lok Woo, and Satnam Dlay. "Statistical Speaker Diarization Using Dependent Combination of Extracted Features." In *2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)*, pp. 291-296. IEEE, 2015 Dec 2, DOI.

[15] Kadhim, Hasan Mohammad-Ali. "Single channel overlapped-speech detection and separation of spontaneous conversations." *PhD diss., Newcastle university*, 2018, URL.

[16] Kanda, Naoyuki, Yashesh Gaur, Xiaofei Wang, Zhong Meng, and Takuya Yoshioka. "Serialized output training for end-to-end overlapped speech recognition." *arXiv preprint arXiv:2002.12687*, 2020, DOI.

[17] Kim, Chanwoo, and Richard M. Stern. "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction." In *tenth annual conference of the international speech communication association*. 2009, 28-31, DOI.

[18] Kim, Chanwoo, and Richard M. Stern. "Power-normalized cepstral coefficients (PNCC) for robust speech recognition." *IEEE/ACM Transactions on audio, speech, and language processing* 24, no. 7, 2016 Mar 23: 1315-1329, DOI.

[19] Landini, Federico, Ondřej Glembek, Pavel Matějka, Johan Rohdin, Lukáš Burget, Mireia Diez, and Anna Silnova. "Analysis of the but diarization system for voxconverse challenge." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5819-5823. IEEE, 2021 Jun 6, DOI.

[20] Martinez-Lucas, Luz, Mohammed Abdelwahab, and Carlos Busso. "The MSP-conversation corpus." *Proc. Interspeech 2020*, 1823-1827, 2020 Oct, DOI.

[21] Park, Tae Jin, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan. "A review of speaker

diarization: Recent advances with deep learning." *Computer Speech & Language* 72, 2022 Mar 1: 101317, DOI.

[22] Singh, Satwinder, Ruili Wang, and Yuanhang Qiu. "DeepF0: End-to-end fundamental frequency estimation for music and speech signals." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 61-65. IEEE, 2021 Jun 6, DOI.

[23] Subakan, Cem, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. "Attention is all you need in speech separation." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21-25. IEEE, 2021 Jun 6, DOI.

[24] Wang, Quan, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno. "Speaker diarization with LSTM." In *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pp. 5239-5243. IEEE, 2018 Apr 15, DOI.

[25] Wang, Yu-Xuan, Jun Du, Maokui He, Shutong Niu, Lei Sun, and Chin-Hui Lee. "Scenario-Dependent Speaker Diarization for DIHARD-III Challenge." In *Interspeech*, pp. 3106-3110. 2021, DOI.

[26] Wong, Eddie, and Sridha Sridharan. "Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification." In *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No. 01EX489)*, pp. 95-98. IEEE, 2001 May 4, DOI.

[27] Zeghidour, Neil, and David Grangier. "Wavesplit: End-to-end speech separation by speaker clustering." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29, 2021 Jul 26: 2840-2849, DOI.

[28] Zhang, Aonan, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang. "Fully supervised speaker diarization." In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6301-6305. IEEE, 2019 May 12, DOI.

Nomenclature

b(t)	Boy speech signal (joule).
$e_b(t)$	Error against the boy signal (joule).
$e_g(t)$	Error against the girl signal (joule).
ERROR	Total time for the reference speaker, wrong speaker attributed (second).
FA	Total time for speaker (hypothesis not a reference speaker attributed (second).
gb(t)	Girl with boy signal (joule).
g(t)	Girl speech signal (joule).
MISS	Total time for reference speaker time, a hypothesis speaker for not attributed (second).
TOTAL	Total reference speech time, which is the sum of the time of the all segments (second).

Subscripts

b	Boy
g	Girl
hyp	Hypothesis
ref	Reference

Abbreviations

Av.	Average value
DER	Diarization Error Rate
DCT	Discrete Cosine Transform
DSP	Digital Signal Processing
DTW	Dynamic Time Warping
LPC	Linear Prediction Cepstral
LPCC	Linear Predictive Coding Coefficients
Max.	Maximum value
MFCC	Mel-Frequency Cepstral Coefficients
Min.	Minimum value
ML	Machine Learning
PDA	Pitch Detection Algorithm
PLPC	Perceptual Linear Prediction Coefficients
PNCC	Power-Normalized Cepstral Coefficients
SAR	Source-to-Artifact Ratio
SDR	Source-to-Distortion Ratio
SIR	Source-to-Interference Ratio
STFT	Short-Time Fourier Transform
TIMIT	Audio and speech dataset of Massachusetts Institute of Technology (MIT)

التعلم الآلي الخاضع للإشراف لتنسيق المتحدث باستخدام معاملات الصوت PNCC-LPCC

***حسن محمدعلي كاظم*[1] *، *علاء حسين أحمد* [2]، *آلاء كريم حسن* [3]، *سعد طه ياسين الفلاحي*[4]**

[1]*قسم الهندسة الكهربائية، كلية الهندسة،جامعة مستنصرية، بغداد، العراق، hasanalmgotir@uomustansiriyah.edu.iq*

[2]*قسم الهندسة الكهربائية،كلية الهندسة،جامعة مستنصرية، بغداد، العراق، alaa75hs@uomustansiriyah.edu.iq*

[3]*قسم الهندسة الكهربائية،كلية الهندسة،جامعة مستنصرية، بغداد، العراق، alaak_eng@uomustansiriyah.edu.iq*

[4]*قسم تقنية الحاسوب،كلية مدينة العلم الجامعة، بغداد، العراق، saad.t.yasin@mauc.edu.iq*

* *الباحث الممثل: حسن محمدعلي كاظم، hasanalmgotir@uomustansiriyah.edu.iq*

نشر في: 30 ايلول 2024

الخلاصة – يوميات المتحدث هي تقنية معالجة إشارات رقمية للكلام، التي تفصل إشارة إدخال واحدة تمثل محادثة غير متقاطعة بين أشخاص إلى إشارات متعددة. تنتمي كل إشارة منفصلة إلى أحد هؤلاء الاشخاص بالإضافة إلى القليل من الخطأ، وهو الكلام الذي ينتمي إلى المتحدثين الآخرين. تنسيق هذا الخطاب عبارة عن حوار لأنهم يتحدثون بشكل غير متزامن. ومن خلال استخدام قاعدة بيانات المتحدثين في هذا البحث، يتم استخراج الميزات الصوتية من الكلام. الاستخراج هو مرحلة التدريب على التعلم الآلي. يمكن لمرحلة التصنيف الثانية بعد ذلك أن تقرر كيفية تقسيم هذه الميزات إلى مجموعات عددها يساوي عدد المتحدثين. يتم استخدام معاملات التنبؤ الخطي (LPCC) ومعاملات Cepstral المعيارية للطاقة (PNCC) بشكل مستقل لإنشاء ميزات الكلام. في هذه البحث، فحص الباحثون الجمع بين ميزات LPCC وPNCC لتشكيل مزيج جديد من الميزات. تساعد المسافة الإقليدية المحسنة مهمة قياس المسافات لتحديد من هو أقرب للمتكلم. نظراً لأن PNCC عبارة عن تحويل غير قابل للعكس، فقد تم أخذ إطار صغير في وسط الإطار الرئيسي الكبير (لأنه يتمتع بالوزن الاكبر) للحصول على إشارات الكلام الأصلية. كان الإجراء فعالاً لتجميع مزيج من إشارتي كلام لأنثى وذكر من مكتبة الصوت القياسية TIMIT، أي نجح في استعادة الكلام الفردي لكل شخص. كان متوسط الاختبارات الموضوعية لمعدل خطأ اليوميات (SDR) للكلام المستعاد 1.8% للإناث، و2.9% للذكور، و2.5% للإناث والذكور بشكل عام. وبالمقارنة مع الأبحاث القياسية الأخرى، كانت التحسينات 6.5% للإناث، و10% للذكور، و8.8% لجميع الإناث والذكور.

الكلمات الرئيسية – تسجيل صوتي للمتحدث، PNCC، LPCC، التجميع، معدل خطأ التسجيل.