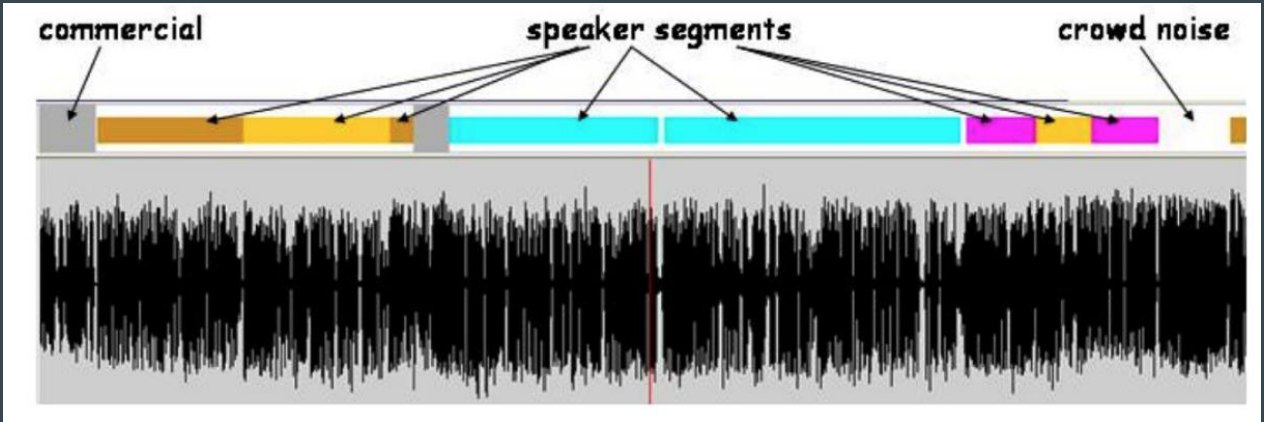
Speech Diarization

Madhu Ashok
&
Erik Núñez

Definition:

The task of determining “who spoke when?” in an audio or video recording that contains an unknown amount of speech and unknown number of Speakers.

1. Identification of each speaker
2. Intervals during which each speaker is active



Problem:

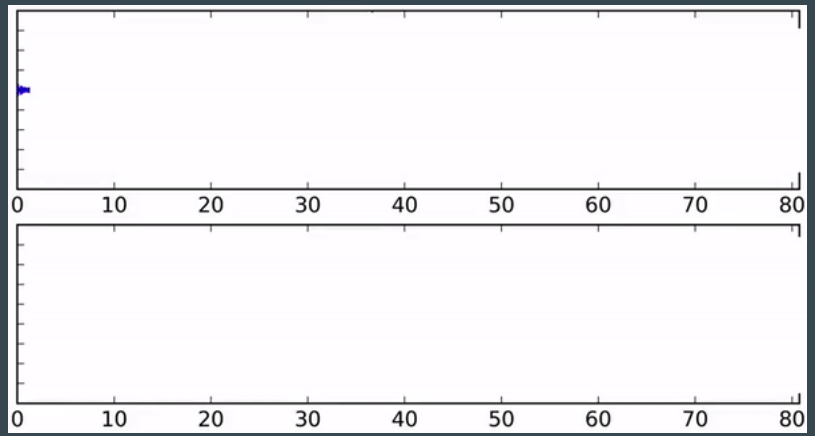
3rd Presidential Debate

Speech diarization systems can be used to alleviate the error introduced to speech recognition systems when noisy data is acquired but transcription is still necessary.

Background

Due to increase in processing power and decrease in storage costs, DARPA Effective, Affordable Reusable Speech-to-text (EARS) at ICSI created the Rich Transcription system back in 2002 as part of a five year project.

RT Transcription is primarily for broadcast news while NIST speaker recognition evaluations are used for telephone speech.

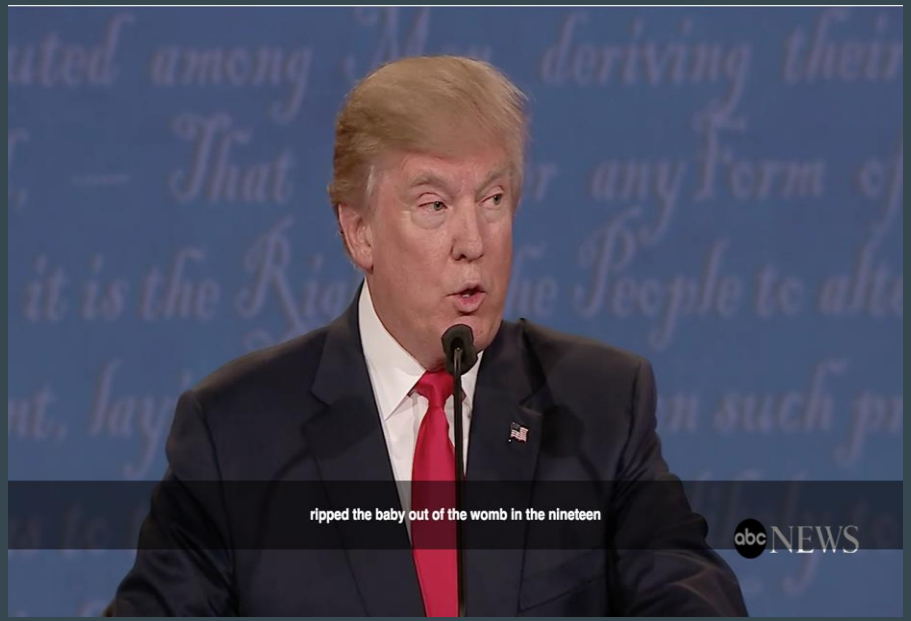


Information Retrieval:

- Broadcast News
- Phone Conversations
- Video Conferencing

Importance/Applications

- Speech and Speaker indexing
- Document Content structuring
- Speaker recognition in the presence of multiple speakers
- Speech to text transcription
- Speech translation
- Rich Transcription (most current technology 2012) sponsored by the NIST



Adapting from Broadcast News to Teleconferencing

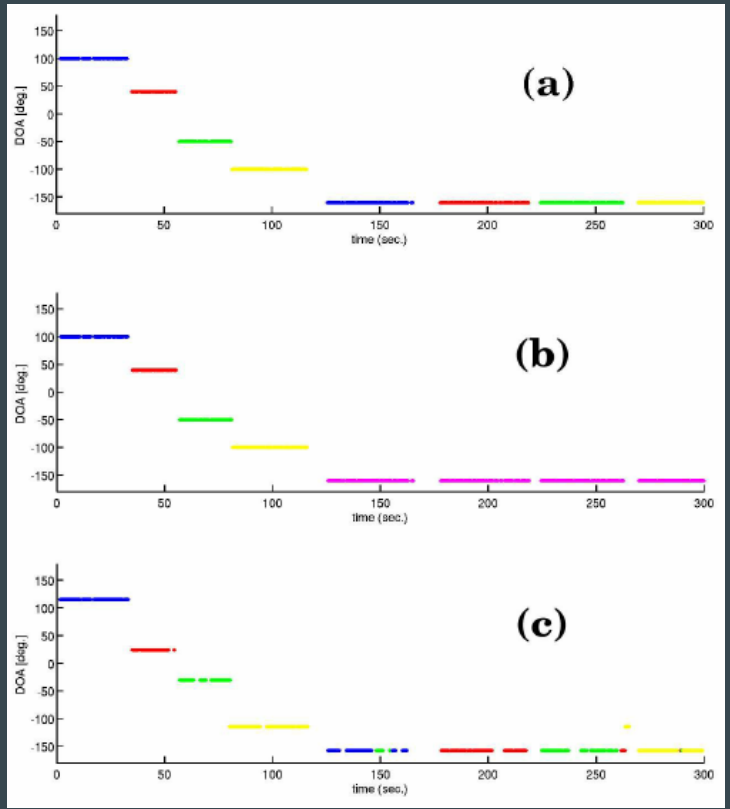
Broadcast news has anchors with close proximity microphones and generally stick to a script. There are also datasets for famous broadcasters that are sufficiently labeled to allow for supervised training.

Higher SNR and lower speaker turn rate in general.

(a) – Ground Truth

(b) – DOA without Speech Diarization

(c) – DOA with Speech Diarization



Adapting from Broadcast News to Teleconferencing

Recordings from desktop or far-field microphones with lower SNR

The difficulty in speaker diarization depends on the amount of prior knowledge allowed, such as example speech from the speakers in the audio can be used to make the task more of a speaker detection task.

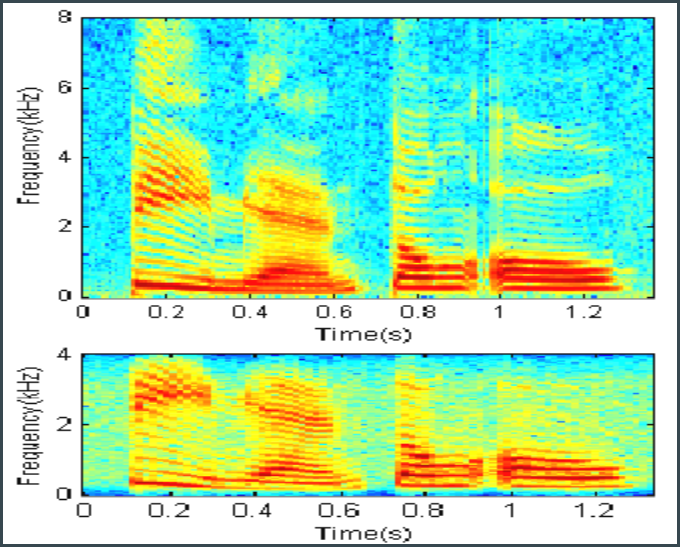
Multiple nearfield microphones are placed near each speaker in an attempt to isolate each source.



Gender/Bandwidth Classification

Speaker segments can become misclassified with gender and introduce error into systems.

Gender classification can yield high accuracy but is computationally expensive if a speech recognition is not available.



Bandwidth classification should be primarily used as a preprocessing method for data as it can be easily determined with low error rates.

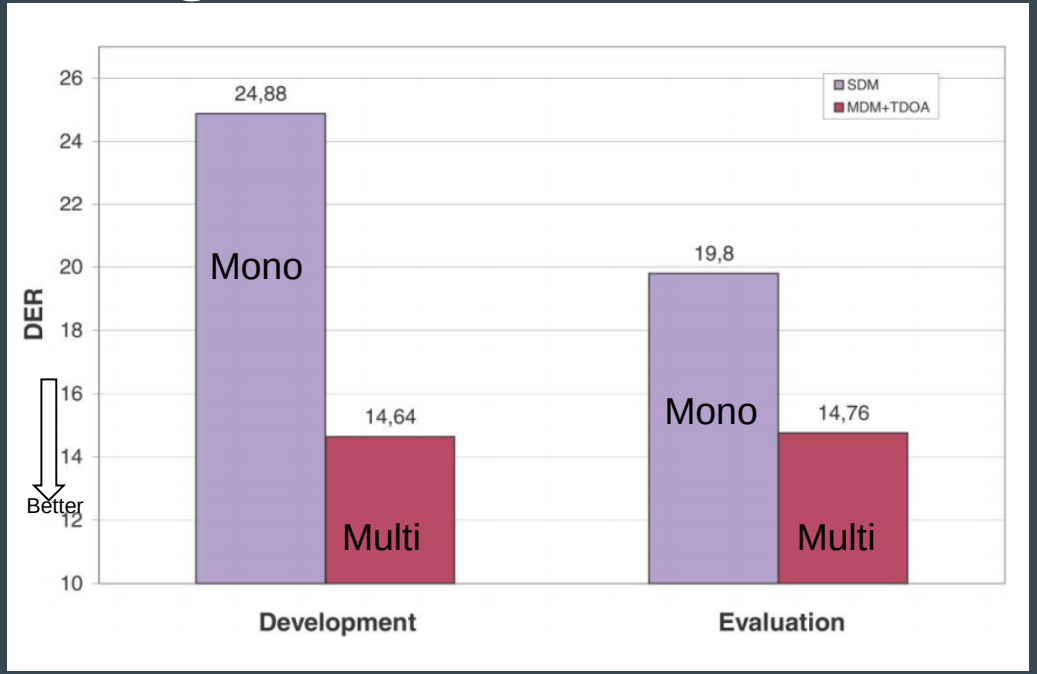
Formants are cut off in telephone conversations due to lower sampling rates.

Algorithms: Acoustic Beamforming

$$y[n] = \sum_{m=1}^{M} W_m[n] x_m[n - \mathrm{TDOA}^{(m,\mathrm{ref})}[n]]$$

W_m - Relative weight for microphone m
$TDOA^{(m, ref)}$ - Time delay of arrival from channel m to the reference
$x_m[n]$ - Signal for each channel, m, at sample n

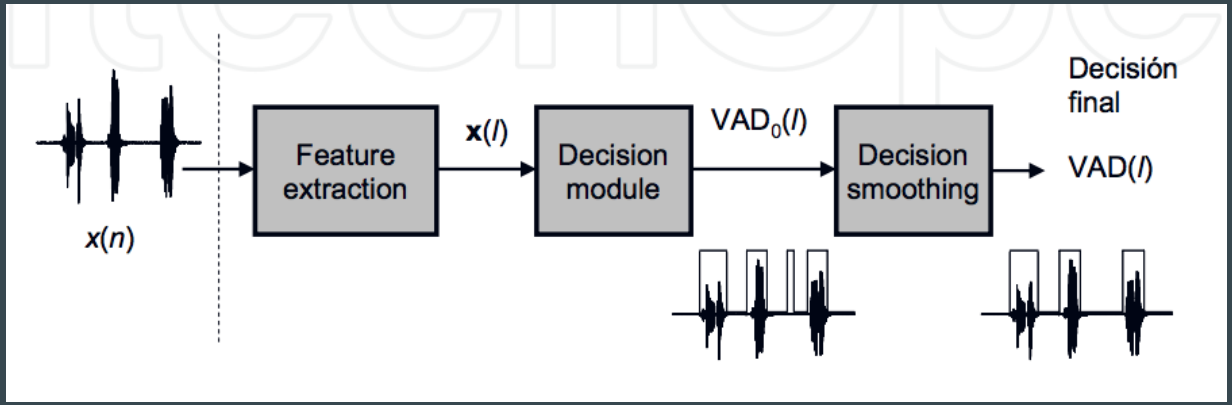
- Find the relative TDOA for each microphone
- Usually unknown locations and gain for each microphone (m)



Diarization Error Rate (DER) using beamforming [X. Anguera]

Algorithms: Speech Activity Detection

- Speech Activity Detection (SAD) - Labeling the speech and non-speech segments (Voice Activity Detection – VAD)
- Poor SAD performance will increase the diarization error rate (DER)
- Feature Extraction - compute discriminative speech features suitable for detection
- Log-Likelihood Ratio (LLR) – activity detector



[J. Ramirez]

Algorithms: Segmentation

- Speaker Segmentation - detecting the change in speaker
- Bayesian Information Criterion (BIC) - sets a penalty term to control missed turns and false detections

$$\Delta BIC = (n_i + n_j)\log|\Sigma| - n_i \log|\Sigma_i| - n_j \log|\Sigma_j| - \lambda\frac{1}{2}(d + \frac{1}{2}d(d+1))\log n \quad (1)$$

$n = n_i + n_j$ - Size of the two clusters

| | - Determinant

d - Dimension of the feature vector space

λ - Weight of the BIC penalty

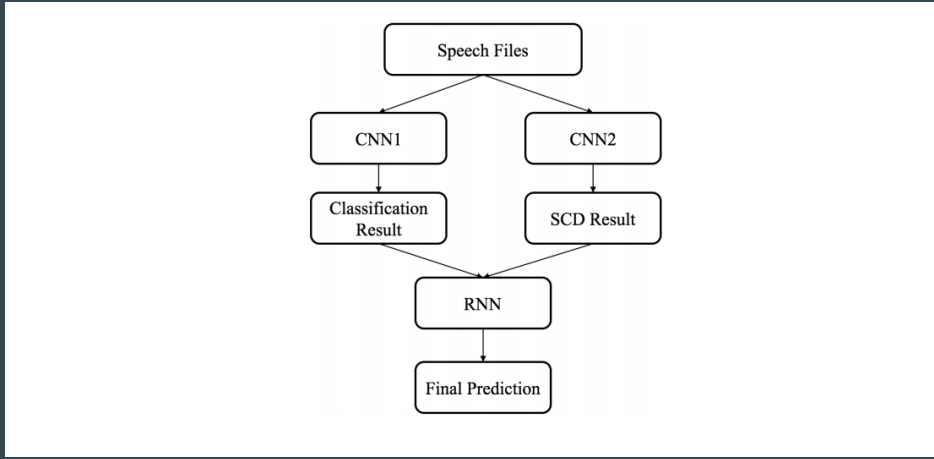
Σ - Covariance matrix estimated on $o_1, o_2, \ldots, o_n$

Σ_i - Covariance matrix estimated on $o_1, o_2, \ldots, o_{ni}$

Σ_j - Covariance matrix estimated on $o_1, o_2, \ldots, o_{nj}$

[Z. Xuan]

Algorithms: Joint Diarization and Clustering with Neural Networks



[Z. Duan]

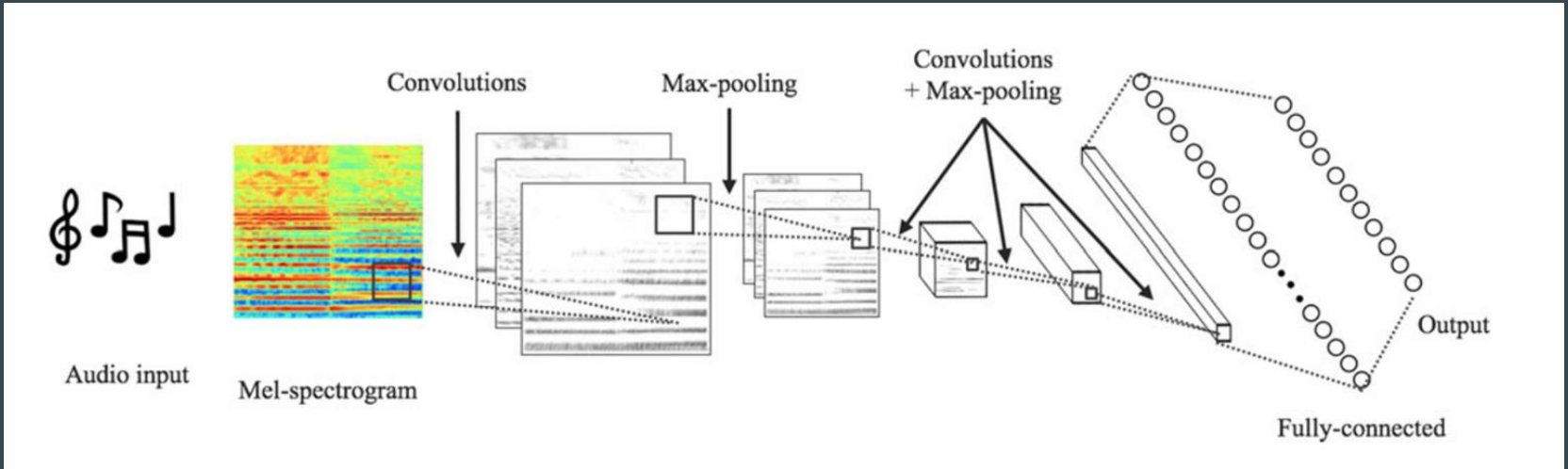
CNN1 - The first convolutional neural network, which identifies the sources

CNN2 - The second convolutional neural network, which classifies when a source changes (Speaker Change Detection)

RNN - Join CNN1 with CNN2 to identify the source and when it is activated

Cont. Neural Networks

- Knowing the number and type of the instrument can significantly improve the performance of audio source separation, automatic music transcription, and genre classification.



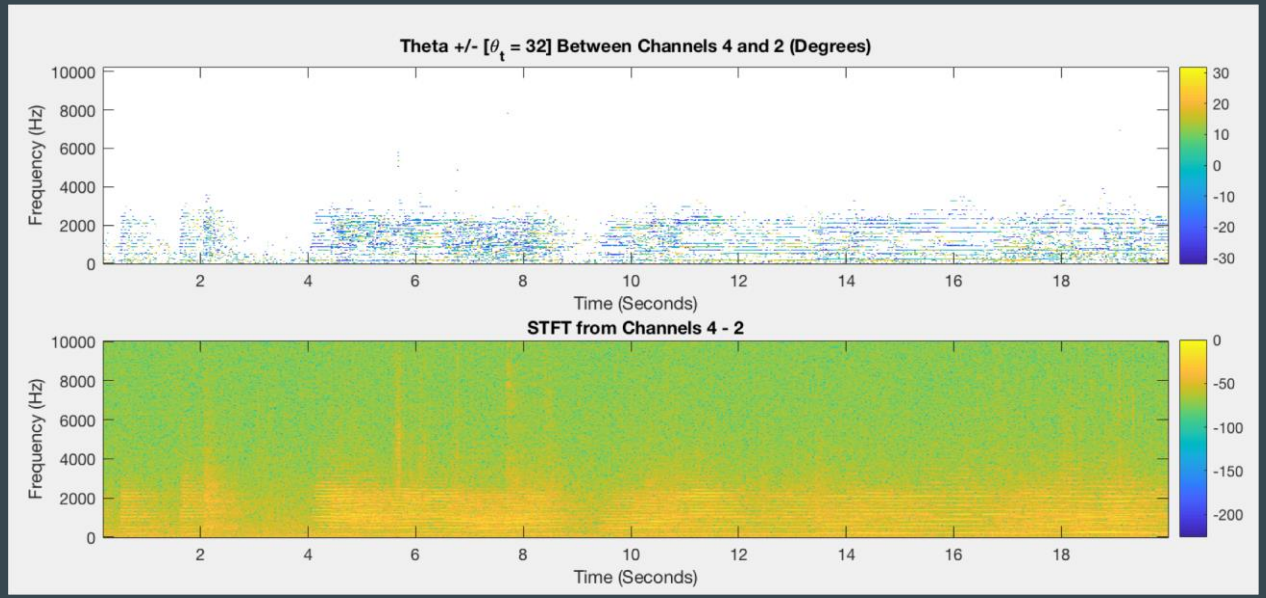
[Y. Han]

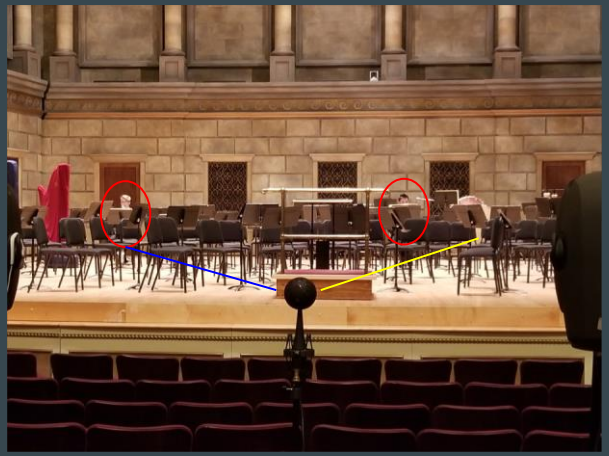
Algorithms: Current Research Directions

- Time-Delay Features
- Use of Prosodic Features in Diarization
- Overlap Detection
- Audiovisual Diarization
- System Combinations

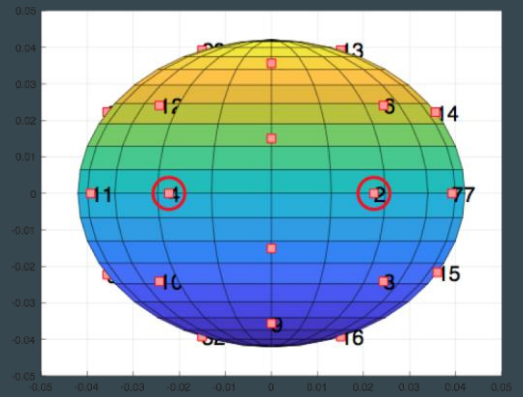
Microphone arrays

Future Research with the 32-Channel Eigenmike





Kodak Hall – Two performers on stage with noise in the audience



View with respect to the stage

Future Work/Limitations

Core component of Automatic Speech Recognition systems, improving accuracy of transcription results.

Tends to be a feature added in later on to improve accuracy results for other systems.

Sources can vary substantially from sampling rates which affect bandwidth,, types of microphones, number and type of non speech sources, number of speakers, duration and sequencing of speaker turns, and any individual spontaneity/style due to the speakers natural speaking style.

References

- K. Church, W. Zhu, J. Vopicka, J. Pelecanos, D. Dimitriadis and P. Fousek, "Speaker diarization: A perspective on challenges and opportunities from theory to practice," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 4950-4954. doi: 10.1109/ICASSP.2017.7953098
- X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland and O. Vinyals, "Speaker Diarization: A Review of Recent Research," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 2, pp. 356-370, Feb. 2012. doi: 10.1109/TASL.2011.2125954
- Y. Han, J. Kim and K. Lee, "Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 1, pp. 208-221, Jan. 2017. doi: 10.1109/TASLP.2016.2632307
- Z. Zhou, Y. Zhang and Z. Duan, "Joint Speaker Diarization and Recognition Using Convolutional and Recurrent Neural Networks," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, 2018, pp. 2496-2500. doi: 10.1109/ICASSP.2018.8461666
- H. Sun, B. Ma, S. Z. K. Khine and H. Li, "Speaker diarization system for RT07 and RT09 meeting room audio," 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, 2010, pp. 4982-4985. doi: 10.1109/ICASSP.2010.5495077
- S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 5, pp. 1557-1565, Sept. 2006. doi: 10.1109/TASL.2006.878256
- Zhang, A., Wang, Q., Zhu, Z., Paisley, J., & Wang, C. (2018). Fully Supervised Speaker Diarization. *arXiv preprint arXiv:1810.04719*.
- X. Anguera, C. Wooters, and J. Hernando, “Acoustic beamforming for speaker diarization of meetings,” IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 7, pp. 2011–2023, Sep. 2007
- J. Ramirez, J. M. Girriz, and J. C. Segura, M. Grimm and K. Kroschel, Eds., “Voice activity detection. Fundamentals and speech recognition system robustness,” in Proc. Robust Speech Recognit. Understand., Vienna, Austria, Jun. 2007, p. 460.
- Zhu, Xuan, et al. “Multi-stage speaker diarization for converence and lecture meetings.” multimodal technologies for perceptions of humans. Springer, Berlin, Heidelberg, 2008. 533-542
- S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada and S. Makino, "A DOA Based Speaker Diarization System for Real Meetings," 2008 Hands-Free Speech Communication and Microphone Arrays, Trento, 2008, pp. 29-32. doi: 10.1109/HSCMA.2008.4538680

Any Questions?

...