dHealth 2019 – From eHealth to dHealth
D. Hayn et al. (Eds.)

doi:10.3233/978-1-61499-971-3-81

Evaluation of Deep Clustering for Diarization of Aphasic Speech

Daniel KLISCHIES[a,1], Christian KOHLSCHEIN[a], Cornelius J. WERNER[b] and Stephan M. JONAS[c]
[a]*Institute of Information Management in Mechanical Engineering, RWTH Aachen University, Germany*
[b]*Department of Neurology, Section Interdisciplinary Geriatrics, University Hospital RWTH Aachen, Germany*
[c]*Department of Informatics, Technical University of Munich, Germany*

Abstract. Speaker attribution and labeling of single channel, multi speaker audio files is an area of active research, since the underlying problems have not been solved satisfactorily yet. This especially holds true for non-standard voices and speech, such as children and impaired speakers. Being able to perform speaker labelling of pathological speech would potentially enable the development of computer assisted diagnosis and treatment systems and is thus a desirable research goal. In this manuscript we investigate on the applicability of embeddings of audio signals, in the form of time and frequency-band based segments, into arbitrary vector spaces on diarization of pathological speech. We focus on modifying an existing embedding estimator such that it can be used for diarization. This is mainly done via clustering the time and frequency band dependant vectors and subsequently performing a majority vote procedure on all frequency dependent vectors of the same time segment to assign a speaker label. The result is evaluated on recordings of interviews of aphasia patients and language therapists. We demonstrate general applicability, with error rates that are close to what has been previously achieved in diarizing children's speech. Additionally, we propose to enhance the processing pipelines with smoothing and a more sophisticated, energy based, voting scheme.

Keywords. diarization, expressive language disorders, machine learning, medical informatics

1. Introduction

Aphasia is a language disorder usually acquired from strokes or other causes of brain damage. It is usually not related to motoric or sensoric incapabilities, but a loss of the brains capability to formulate language. The gold standard for aphasia classification and severity measurement in Germany is the Aachen aphasia test (AAT) [1]. It consists of several procedures, testing the patient's linguistic capabilities in scenarios like image description, storytelling and spontaneous speech. Conducting and evaluating a complete AAT takes up to eight hours of work by a professional speech and language therapist or neurologist and requires that the patient is present at the clinic.

[1] Corresponding Author: Daniel Klischies, Institute of Information Management in Mechanical Engineering (IMA), RWTH Aachen University, Germany, E-Mail: daniel.klischies@ima-ifu.rwth-aachen.de.

Our long term goal is to develop a method to automatically estimate aphasia severity and syndrome classification based on a preexisting recording of a patient interview. In order to do this, we need to separate the therapist's speech segments from the patient's speech segments. This process, also called diarization, has seen significant research interest in the past. While originally conducted using clustering procedures based on immediate features of the underlying audio signal, recent developments (such as [2, 3]) suggested that generating clustering features using neural networks yields better results.

The specific use case of aphasic speech also yields some additional challenges and characteristics with regard to diarization. Since a symptom of aphasia can be stuttering and extensive use of filler words, we require that these are included in the diarization output. We also cannot rely the diarization on semantic or syntactic linguistic properties, since both capabilities might be severely reduced as an effect of aphasia. Lastly, recordings of aphasia patients are rare and hard to obtain, because the prevalence of aphasia within the population is relatively low and obtaining recordings usually includes adhering to strict data protection rules. This results in significant problems when training any machine learning based classifier, as training material is rare. Additionally, virtually all recordings of aphasia patients are single source recordings, eliminating the possibility to use multi-source diarization procedures. This specifically holds true for a set of aphasic speech data we currently possess and are planning to analyze, and which was the original motivation to perform single source diarization.

In general, single source speaker diarization systems require a set of metrics that can be used to locally cluster temporal segments of speech, such that each cluster represents a segment of speech by the same speaker. This can be implemented either bottom up, by first creating many small segments and subsequently merging those segments, or top down by splitting segments as long as they are suspected to be comprised of more than one speaker. One possibility to perform bottom up clustering is to implement the initial splitting based on a sliding window over the audio signal and subsequently clustering these segments. Such an approach has recently been investigated by Wang et al. [3]. Their diarization system uses a long short-term memory (LSTM) network, which is a memory cell based recurrent neural network [4], to derive a vector space embedding of the sliding window segments and clusters them using different procedures. The most promising clustering procedures are k-means clustering and spectral clustering, with a diarization error rate of roughly 12%, depending on the evaluation data set. They compared their results to the diarization error rate of a similar system that uses Gaussian mixture models (GMMs [5, pp. 40-42]) instead of LSTMs to derive the embeddings, which performed worse by at least 8 percentage points.

A combination of both, bottom up and top down clustering, has been implemented by Bozonnet et al. in [6]. Their integrated approach led to an improvement in diarization error rate by 4 percentage points, although generally performing worse than the system developed by Wang et al, albeit their usage of different evaluation data. This is most probably because the latter uses LSTM based embeddings, while the system by Bozonnet used Gaussian mixture models.

In 2010, Meignier and Merlin published a paper describing the LIUM toolkit for development of diarization systems [7]. Contrary to the other systems presented, this system provides several building blocks for the implementation of diarization utilities, based on agglomerative clustering. Due to its release date, this framework does not use LSTMs but the older GMM approach.

Lastly, in 2016 Hershey et al. proposed a method to generate vector space embeddings from speech data using LSTMs [2, 4]. These embeddings are generated based on logarithmically scaled short term Fourier transformations of segments originating from the input speech audio file. Each of these segments is comprised of a time and frequency interval of said input audio. The LSTM is subsequently used to compute and optimize an affinity matrix, denoting which time frequency segments belong to the same speaker. Subsequently, several clustering methods such as the aforementioned spectral clustering and k-means clustering are used to determine which segments belong to the same speaker. Since the clustering is not only time but also frequency band agnostic, this allows to separate overlapping speech, which is an inherently more complex task.

Here, we investigate on how to adopt the method proposed by Hershey et al. to speaker diarization, specifically for diarizing aphasic speech.

2. Methods

Applying deep clustering as introduced by Hershey et al. in [2] in practice leads to several details that might influence the results dramatically. These details resolve around how much training data is required to train an estimator, such that the embeddings are sufficiently discriminative to fulfill the aforementioned criteria for successful clusterings. Additionally, we are not interested in speech separation but diarization, requiring a slight alteration of Hershey's proposed algorithm. In order to collapse the time frequency bins into time bins, we employ a majority voting scheme: For each set of time frequency bins representing the same time slot t, we count how many frequencies have been assigned to which speaker. In the next step we assign the time slot t of the input signal to the speaker to whom the most time frequency bins were assigned. Under the hypothesis that the acoustically dominant speaker of a time slot also dominates most of the time frequency bins of said time slot, this majority voting allows us to diarize an input file such that we always get the whole spectrum assigned to a single speaker. This saves us from having to deal with a signal reconstruction problem that the original separation procedure suffers from: If one does assign time-frequency bins of a signal to different speakers, all those frequency bands that have not been assigned to a speaker would be missing from the output signal. Isik et al. proposed some reconstruction methods for these parts of the signal [8], but since we are dealing with pathological speech any reconstruction methods based on assumptions of non-pathological speech could introduce incorrect additions and reduce the quality of a diagnosis based on the reconstructed signal.

Our implementation of the deep clustering algorithm itself is based on an implementation by Haroran Zhou, who implemented deep clustering using Tensorflow (https://github.com/zhr1201/deep-clustering). We modified his work, such that it supports Python 3, resolved some minor bugs and adopted the frequency band majority voting strategy presented in the previous section.

We preprocess data by down-sampling the signal to 8kHz, and generate 129 Fourier transformation points per frame, with a window size of 256 separate Fourier transformations, such that we get Hanning windows of length 256/8000Hz = 0.032s. The network itself consists of four bidirectional LSTM layers with 300 memory units per layer, followed by a layer with a hyperbolic tangent activation function to estimate the embedding. Finally, the embedding is normalized based on its L2 norm.

We use a dropout of 50% for the forward propagation and 20% for the recurrent propagation of errors, the estimated embedding space has 40 dimensions.

The classifier has been trained for a week using an Nvidia Titan X (Pascal architecture), which was sufficient for 352000 training steps. For the training corpus, we mixed (non-aphasic) speech files from the 360 hours LibriSpeech audio book corpus [9], such that we get training files with two simultaneously speaking speakers per file. We do this by combining 20 files per speaker with some other randomly chosen file containing another speaker. The training is thus based on the original usage of the classifier, as proposed by Hershey et al. and the resulting classifier could also be used for speech separation. Our majority voting scheme is only applied after the training is completed and the classifier is being evaluated.

2.1. Diarization error rate

In order to evaluate the results, we use a slightly modified version of the diarization error rate (DER), which was originally proposed by the National Institute of Standards and Technology (NIST) (cf. [10]). Given a prediction and a ground truth set of speaker labels, the DER quantifies the correctness of the prediction. The ultimate goal is to develop a diarization procedure that yields predictions with a DER of 0. While the NIST definition measures how much of the overall recording time was incorrectly attributed, we want to measure how many of the potential speaker labels are incorrect. This penalizes classifiers that do not detect overlapping speech properly more than the NIST definition: In the NIST definition, a segment that actually contains two speakers but was classified as silence increases the DER just as much as a segment that contains two speakers but was classified to contain one speaker. In our definition, classifying this segment to contain silence is twice as bad as classifying it to contain a single speaker.

For an audio recording of length T_Σ with a frame rate B, for which we know that it contains N speakers, we define the maximum amount of possibly incorrectly assigned labels $\mathrm{E}_{\mathrm{max}} = \mathrm{T}_\Sigma \cdot B \cdot N$. Furthermore, we define $L \in \mathrm{F}_2^{(\mathrm{T}_\Sigma \cdot \mathrm{B}) \times 2}$ to be our ground truth speaker label matrix where $L_{i,j} = 1$ iff in the i^{th} frame, the j^{th} speaker is active, and $P \in \mathrm{F}_2^{(\mathrm{T}_\Sigma \cdot \mathrm{B}) \times 2}$ to be the estimated speaker label matrix. Then we can decompose the DER of P given L into the following components:

$$E_{fa} = \sum_{\substack{0 \leq i < \mathrm{T}_\Sigma \cdot B \\ \wedge L_{i,\cdot} = 0}} \frac{\sum_{j=0}^{N} P_{i,j}}{E_{max}} \tag{1}$$

$$E_{miss} = \sum_{\substack{0 \leq i < \mathrm{T}_\Sigma \cdot B \\ \wedge P_{i,\cdot} = 0}} \frac{\sum_{j=0}^{N} L_{i,j}}{E_{max}} \tag{2}$$

$$E_{error} = \sum_{\substack{0 \leq i < \mathrm{T}_\Sigma \cdot B \\ \wedge L_{i,\cdot} = 0 \, \wedge 0 < L_{i,\cdot}}} \frac{\sum_{j=0}^{N} |(L_{i,\cdot} - P_{i,\cdot})_j|}{E_{max}} \tag{3}$$

E_{fa} is the false alarm rate, which we define as the percentage of possible speech label tags that were marked as non-silence but were actually silence. A high false alarm rate indicates that there is an issue with the voice activity detector (VAD) of the diarization procedure. Analogously E_{miss} is the percentage of speech labels that were classified as silence but were actually speech. If this value is high, then the VAD algorithm is too restrictive, as it misses some speech. Lastly E_{error} is the percentage of incorrectly assigned speech labels, in regions where there was no silence according to the ground truth and the diarization algorithm's VAD.

Given these components, the DER can be calculated as described in equation 4.

$$E_{DER} = E_{fa} + E_{miss} + E_{error} \tag{4}$$

2.2. Benchmarking

We developed an automatic benchmarking procedure along with a set of Python scripts performing the benchmarks. These scripts take a list of input folders containing audio files, where each folder represents a speaker. Each file in each folder is then mixed alternatingly with every file of the other speakers. Information about which part belonged to which speaker is stored, in order to later on compare this ground truth to the results of the diarization. Therefore, we evaluate the quality of deep clustering on a computer-generated test set. This allows us to generate many different test audio files with different lengths of individual utterances and the whole file, allowing more flexible test scenarios compared to using the original recordings. Additionally, this allows us to evaluate the performance of diarization algorithms on arbitrary speaker pairs, not just those that are present in the test data set.

Currently our benchmarking script supports the following options:

1. Arbitrary amount of evaluation speakers, but a fixed amount of two speakers per generated output file
2. Minimum/maximum length of each utterance in the output file
3. Length of a crossfaded overlap between utterances
4. Length of silence between utterances
5. Maximum number of output files to generate
6. Minimum length of output file (if two input files do not result in sufficient length of the output file, the algorithm will append additional files from the same speakers)

The scripts are capable of benchmarking arbitrary diarization algorithms and implementations. Each of these implementations has to provide a Python 3 class with a method, that takes a file path to a sample audio file to diarize, and optionally an integer specifying into how many prediction frames this file should be split. Speaker vectors generated by this method must always be two dimensional, and the overall return value of the method must follow the same semantics that we defined for the prediction matrix P in our definitions of DER (see equation 4 in section 2.1).

After the diarization procedure returned its predicted diarization P to the benchmarking script, it compares this value to the ground truth labels L that are based on its knowledge about the original combination of different speakers and quantifies this using the DER metric presented in the previous subsection 2.1. Lastly, the results are summarized and mean, minimum, maximum, standard deviation and variance are calculated over the set of diarization error rates.

3. Results

Our evaluation data set is based on the AphasiaBank dataset [11]. AphasiaBank is a data set composed of transcribed video recordings of semi standardized interview scenarios between aphasia patients and therapists. We downloaded and automatically split all those recordings into utterances labeled with information whether the therapist or the patient is speaking in that particular utterance, based on the timestamps and speaker labels of the transcripts. Since the therapist usually does not change between different recordings of the same data set, we store all therapist utterances of the same institution as it came from one recording, while patient utterances are separated such that each recording leads to a separate set of patient utterances.

We recombined a subset of the AphasiaBank speaker files, such that we get audio files with a minimum length of 5 seconds and at most 3 seconds per utterance. The latter value roughly matches average speaker durations in common evaluation data sets for diarization of healthy speech [12]. For each speaker of the subset, we randomly choose at least 5 utterances and combined each of them with an utterance from another, randomly chosen speaker. If that combination was not at least 5 seconds long, we appended additional utterances from the same speakers until 5 seconds of file length were reached. This length requirements ensures that we get a balanced set of evaluation data. This is particularly relevant because, depending on the aphasia syndrome, patients tend to speak significantly longer or shorter than the therapist. Additionally, we did not only mix speech of patients with speech of therapists, but also with speech of other patients. This allows us to judge the quality of the classifier for diarization of aphasic speech in general, and not only in scenarios where exactly one speaker suffers from aphasia. This would not be possible, if we would not have recombined the files, as we do not possess recordings containing multiple patients.

The result of this process were 125 separate audio files. Diarizing them with deep clustering led to a DER of 27.94%, minimally 13.9%, maximally 39.19% and a standard deviation of 0.0439. Since the way we compose the input file does not allow for overlap (no crossfade) or gaps, the "false alarm" and "miss" error rates do not play a role in this evaluation, and we only rely on the "error" part of the diarization metric.

4. Discussion and Future Work

We conducted an evaluation using specifically remixed audio segments of the AphasiaBank data set, ensuring a well-balanced amount of speaker duration in all evaluation files.

Comparing the achieved result to other diarization systems is hardly possible, due to the fact that most speaker diarization systems are evaluated on professionally recorded, healthy adult speech. An evaluation conducted by Anguera et al. in 2012 showed diarization error rates between 7% and 49% for the RT09 data set, consisting of healthy speech [12]. However, large amounts of this error are caused by overlapping speech, which is a scenario we have not evaluated yet. Incorrectly attributed speaker information accounts for 5 to 10 percentage points of the diarization error rate in this scenario. While this is considerably better than our implementation of deep clustering, it was achieved on healthy speech that is probably an easier task, and on recordings that were made using professional equipment.

In 2018 Cristia et al. attempted to diarize conversational speech of children [13]. Since speech of children is located on a much smaller and therefore less discriminative frequency band than adult speech, they argue that this is a much harder problem then diarization of adult speech. Furthermore, since children are still learning to comprehend sentences and are generally more affected by language disorders [14], diarization of such speech might be more comparable to aphasic speech. Their system led to a DER of 20% to 40% for two speakers.

The diarization result turned out to be mediocre compared to diarization systems for non-aphasic speech. This could be related to insufficient amount of training data, improvable network architecture or an incorrect voting scheme. Compared to common diarization systems for healthy speech, our approach is worse by 10 to 15 percentage points but performs roughly on the same level as diarization of children's speech, which is also prone to some issues that impair diarization of aphasic speech.

The performance of the deep clustering algorithm leaves a lot of room for improvement. Some of these are general improvements that are beneficial for both, pathologic and healthy speech diarization tasks, including improvements to the voice activity detector by means such as adding a dynamic threshold to determine whether the signal energy is sufficiently to be speech. More advanced techniques like spectral analysis, as proposed by Ma et al. in [15] could improve this even more and might even be able to cope with significant background noise.

The majority voting scheme that we used to adopt the algorithm to diarization instead of separation might also be improvable. In its current implementation, a speaker who gets assigned many low energy frequency bands of a time slot wins over a speaker to whom few, but high energy bands were assigned. Taking the energy of the frequency bands into account would lead to a weighted voting scheme that might work better than our naive implementation.

The exact parameterization of the neural network, along with the training data used, is currently also mostly arbitrary. Investigating how different parameters affect the diarization error rate of the final classifier could lead to the ability to fine tune the system to pathologic and especially aphasic speech. Apart from hyper parameter tuning, that could be performed using grid search, it would be interesting to optimize the training data set. Due to the lack of sufficient data, it is highly unlikely that it will become possible to train the classifier exclusively on aphasic speech. It is therefore desirable to determine a way to apply transfer learning to the classifier, such that it is trained on healthy speech and subsequently adapted to pathologic speech in a way that does not require huge amounts of pathologic speech. Unfortunately, any changes to the neural network's layout require re-training the classifier, which requires a significant amount of computing time.

References

[1] Walter Huber et al., *Aachener Aphasie Test (AAT): Handanweisung*, Verlag für Psychologie, Hogrefe, 1983.

[2] John R. Hershey et al., Deep clustering: Discriminative embeddings for segmentation and separation, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016), 31-35.

[3] Quan Wang et al., Speaker Diarization with LSTM, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), 5239-5243.

[4] Sepp Hochreiter and Jürgen Schmidhuber, Long Short-Term Memory, *Neural Computation* 9.8 (1997), 1735-1780.

[5] Geoffrey McLachlan and David Peel, *Finite Mixture Models*, John Wiley & Sons, Hoboken, 2000.

[6] Simon Bozonnet et al., An integrated top-down/bottom-up approach to speaker diarization, *Eleventh Annual Conference of the International Speech Communication Association (INTERSPEECH)* (2010), 2646-2649.

[7] Sylvain Meignier and Teva Merlin, *LIUM SpkDiarization: an open source toolkit for diarization*, CMU SPUD Workshop, 2010.

[8] Yusuf Isik et al., Single-channel multi-speaker separation using deep clustering, *17th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (2016), 545-549.

[9] Vassil Panayotov et al., Librispeech: an ASR corpus based on public domain audio books, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), 5206-5210.

[10] Jonathan G. Fiscus et al., *The rich transcription 2005 spring meeting recognition evaluation*, in: International Workshop on Machine Learning for Multimodal Interaction, Springer, Heidelberg, 2005, 369-389.

[11] Brian MacWhinney et al., AphasiaBank: Methods for studying discourse, *Aphasiology* 25.11 (2011), 1286-1307.

[12] Xavier Anguera et al., Speaker diarization: A review of recent research, *IEEE Transactions on Audio, Speech and Language Processing* 20.2 (2012), 356-370.

[13] Alejandrina Cristia et al., Talker diarization in the wild: The case of child-centered daylong audio-recordings, *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (2018), 2583-2587.

[14] Victoria M. Garlock et al., Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults, *Journal of Memory and language* 45.3 (2001), 468-492.

[15] Yanna Ma and Akinori Nishihara, Efficient voice activity detection algorithm using long-term spectral flatness measure, *EURASIP Journal on Audio, Speech, and Music Processing 2013.1* (2013), 87.