A Free Synthetic Corpus for Speaker Diarization Research

Erik Edwards[1], Michael Brenndoerfer[2], Amanda Robinson[1], Najmeh Sadoughi[1], Greg P. Finley[1], Maxim Korenevsky[1], Nico Axtmann[3], Mark Miller[1], and David Suendermann-Oeft[1]

[1] EMR.AI Inc., San Francisco, CA, USA
[2] University of California Berkeley, CA, USA
[3] DHBW, Karlsruhe, Germany
erik.edwards@emr.ai

Abstract. A synthetic corpus of dialogs was constructed from the LibriSpeech corpus, and is made freely available for diarization research. It includes over 90 hours of training data, and over 9 hours each of development and test data. Both 2-person and 3-person dialogs, with and without overlap, are included. Timing information is provided in several formats, and includes not only speaker segmentations, but also phoneme segmentations. As such, it is a useful starting point for general, particularly early-stage, diarization system development.

Keywords: speaker diarization · speech activity detection · open-source corpora

1 Introduction

1.1 Background and motivation

Speaker diarization is the task of segmenting an audio file with multiple speakers into speaker turns, also known as “speaker indexing” or the “who spoke when” question. This task was first considered for air-traffic control recordings [13,30,34,38], and has since been applied to a variety of applications [1,2,25], most often to 2-person telephone conversations [8,36,24], broadcast radio and television [12,33], and many-person (e.g. 4-10+) meetings [43,4]. Our own application is doctor-patient dialogs [9], usually consisting of 2 speakers, but occasionally 3 speakers, and only very rarely 4+ speakers. We were not able to identify a suitable training corpus for diarization system development, which is understandable given that medical dialogs contain sensitive personal information. A recently-released diarization challenge set (for the “DIHARD” challenge) included some clinical interviews with doctors and autistic children, but it was required to delete the data following the challenge. Also, the speech of children may not be considered to be a typical case study for general system development. Other data sets are proprietary and seem particular to a given recording channel and/or background noise condition (e.g. air-traffic control). These do not seem ideal for our application or for general system development, where one might prefer to obtain clean speech and then corrupt it with background noise suitable to the application [21,46]. We decided therefore to make our own synthetic corpus of dialogs, which we make freely available for general use, particularly for early-stage and general diarization system development. Of course, this is not intended to replace real-world data, and each applied worker must also obtain data from their own domain.

The earliest approaches to diarization used a “bottom-up” approach of clustering feature vectors by similarity [13,30]. These are also called “unsupervised” in the sense that they require no labeled training data [34]. Although these approaches have remained heavily used in the literature [4,2], later systems began to introduce “top-down” or “supervised” approaches [38,43,12]. These require a fair amount of labeled training data in addition to test data. In fact, the first such top-down study [38] was also the first to introduce synthetic dialog data for training purposes. Recent diarization approaches utilize neural networks [23,18,20,45,41], and these can likewise require a large amount of training data. However, the manual segmentation of dialog data is remarkably difficult and time-consuming (as we have attempted ourselves), and therefore prohibitive for most groups undertaking to get started with system development. Moreover, to avoid over-tuning to the test set during system development and architecture search, it is strongly preferable to have separate development and test data sets.

A final motivation for our synthetic corpus is that we desired to study the issue of “phoneme specificity” or “phone adaptive training” in speaker diarization [17,7,42,47,5,31,35,44]. This refers to the fact that phoneme acoustic differences confound the detection of speaker acoustic differences. That is, for example, the fricatives of two speakers may be more similar than the fricatives and vowels of the same speaker. In order to address this issue, one generally requires a corpus wherein the phone identities and segmentations are available. We introduce such a corpus here, by using methodology from automatic speech recognition (ASR) to obtain forced alignments of phoneme labels.

1.2 Brief review of diarization data sets

The first diarization data studied was air-traffic control recordings [13,30,34,38], and an early study of a 5-person meeting quickly followed [43]. The 1997 DARPA Speech Recognition Workshop introduced the ARPA Hub4 task, to transcribe radio and television broadcasts [12,33]. This was the first in a series of diarization and related tasks from ARPA (Advanced Research Project Agency) and NIST (National Institute of Standards and Technology), and over 100 publications have been dedicated to the diarization of such broadcasts. We have not been able to locate the past NIST data sets, and recent ones appear to be accessible only with an LDC (Linguistic Data Consortium) account. Also, they can contain music or other background noises, and they do not generally include a large training set or phonemic information. The second major domain of diarization research (also over 100 publications) has been multi-person meetings, particularly following the introduction of widely-used corpora of meeting data, namely the ISL Meeting Corpus [6], the ICSI Meeting corpus [19], the AMI corpus [15], and various meeting data sets from NIST, e.g. [11]. Although these are excellent for their domain of application, they involve many speakers (at least 3 speakers, and 4+ speakers in the great majority) and again a particular audio-channel/background-noise scenario. This may not be suitable for early-stage or general diarization system development, or for research focused on 2-3 speakers. Of these, only the AMI corpus (involving 4+ speakers of British or European English) is freely available with a liberal usage license. The third major domain of diarization research has concerned 2-person telephone conversations, of which the stand-out data set has been the Switchboard corpus [14]. This is by far the closest data set to our intended application, but it also has a few drawbacks: It is only available via LDC account, it is sampled at 8 kHz, it seems particular to the given audio channel, and exact overlapped-speech information may not be obtainable. Therefore, it was deemed that, for general, open-source use, particularly outside of the three major application domains, a free synthetic diarization corpus would be necessary, and likely useful to others as well.

We therefore focused on finding previous synthetic diarization corpora. As mentioned above, the first to introduce synthetic dialog data [38] was also the first top-down study, where availability of training data becomes critical. Another early top-down study [39] likewise used a simulated dialog corpus, for which they cited a CD-ROM. Neither of these early synthetic corpora are currently available to our knowledge. Almost no mention of synthetic data was made in the years following the 1997 NIST set. We find exactly 2 artificial conversations made from TIMIT data [8,22,40], a small synthetic test set from TIMIT data [10], and one large synthetic set made from TIMIT [26]. The later was only described in a few sentences, but appears quite similar in motivation to ours (e.g., conversations of 2-6 speakers). Unfortunately, none of these TIMIT-based sets are available to our knowledge. A set of synthetic Spanish conversations was found [3], but we do not consider non-English sets here.

Therefore, we have developed our own synthetic corpus as a basic starting point for diarization research, derived from the freely available and open-source LibriSpeech corpus [28]. This synthetic diarization corpus is freely available for download at:

https://github.com/EMRAI/emrai-synthetic-diarization-corpus.

2 Synthetic diarization corpus

The LibriSpeech corpus consists of sections of English audio books recorded at 16 kHz sample rate [28], usually with clear articulation and high-quality audio. It was expected therefore that forced alignment could produce highly accurate (albeit not perfect) phonemic segmentations. The open-source and widely-used Kaldi speech recognition toolkit [29] includes a recipe for ASR training and alignment of the LibriSpeech corpus. The use of this ASR set is also advantageous because some analyses from the ASR pipeline can be used in diarization. For example, if a universal background model (UBM) or i-vector extractor is trained on the LibriSpeech ASR corpus, it could be used on the synthetic diarization data as well.

In brief outline, we have constructed our synthetic corpus as follows (further details will be available from the download page of the corpus). For training data, we use the “train clean 100” subset of the LibriSpeech corpus with 100.6 hours of audio. This consists of 585 chapters read by 251 unique speakers (126 male, 125 female), where each chapter has up to 129 utterances. We ranked chapters according to number of utterances, and discarded chapters with fewer than 4 utterances. Alternating chapters in this ranked list were combined into 2-speaker dialogs, with care not to combine the same speaker into a single dialog. The utterances from the 2 speakers were simply alternated until one of the 2 speakers had no further utterances. This resulted in dialogs with 13-259 utterances (median 84). Speakers were combined without respect to gender, resulting in 73 female-female, 65 male-male, and 154 female-male dialogs (292 dialogs total). Dialogs ranged in duration from 2.7-49.6 min (median 17.5 min), yielding 98.15 hours in the total training corpus. The LibriSpeech “dev clean”, “dev other”, “test clean”, and “test other” sets were likewise prepared for diarization development and test sets (Table 1).

Table 1. Synthetic 2-person corpus with no overlap

	Dialogs	Utts (Turns)	Tokens	Hours
Train	292	28522	989715	98.15
Dev-clean	48	2673	53765	4.98
Dev-other	45	2822	50227	4.69
Test-clean	43	2605	52279	5.07
Test-other	45	2861	51305	4.85

Inspired by published statistics of natural conversations [37,16], a small random gap was inserted between speaker turns, as sampled from a Rayleigh distribution with scale parameter (mode) of 200 ms. The longest random draws (i.e. from the long tail of the Rayleigh distribution) were discarded, given that gaps in natural conversations are bounded to some finite value. The actually-used samples ranged from 2 to 819 ms with a mean gap of 240 ms. In each original audio file, the leading/trailing silences were tapered linearly to 0 at start/end, such that no audible transient occurs between speaker turns (i.e. the silent portions transition smoothly into each other). Successive wav files were linearly added into the dialog waveform, with the appropriate offsets, and checked so that no sample accidentally exceeded a range of ±1.

Timing information is provided in 3 formats: 1) the Kaldi .ctm format; 2) the NIST .rttm format [27], as required by the widely-used md-eval-v21.pl script for computing the diarization error rate (DER) [1]; and 3) a simple frame-by-frame list of integer labels. In the later, 0 indicates silence, 1 indicates speaker 1, and 2 indicates speaker 2, etc. Integers greater than 10 indicate overlap. In case the direction of overlap is important, these are coded such that "12" means overlap as speaker 1 transitions to speaker 2, and "21" means overlap as speaker 2 transitions to speaker 1. But if the user is only interested in "overlap", then all integers greater than 10 can be collapsed into one category.

For the NIST .rttm format, we provide two versions. In the first, only speaker turns are indicated (with labels 1, 2, etc.), and where all within-speaker gaps of less than 200 ms are ignored, i.e. labeled as speech. This appears to be the most widely used threshold currently, whereas a previous standard used a threshold of 300 ms [27]. In the second set of .rttm files provided, all silences, including gaps less than 200 ms, are explicitly included (with label 0). From these, users could make other thresholds of within-speaker gaps to ignore.

The dialog .ctm files include the timing information for individual phonemes, as obtained by forced alignment (from the tri4b stage of the Kaldi recipe for the LibriSpeech ASR corpus, using the Kaldi "ali2phones" utility [29]). These .ctm files from the original forced alignments were simply mapped to the new timeline of the dialog. We followed the provided standard recipe for the ASR pipeline, except that we used our own lexicon, for reasons that will be presented in a separate contribution. In brief, we have been studying a syllabic approach to ASR, and have developed a lexicon with syllabic phonology for these purposes. This has resulted in $\sim 20\,\%$ relative improvement in WER, and so this was preferred for forced alignments as well. Moreover, we sought to investigate the use of syllabic structure in diarization (see companion paper), which requires syllabic information from the alignments. Our expanded phone set can be mapped back to the usual ARPAbet phones [32], if desired. Since forced alignment does not work for out-of-vocabulary words, we manually added all such words to our lexicon. This is one of the reasons that we use only the 100-hour "train clean" subset of the full LibriSpeech training data.

A second version of the corpus incorporates speaker overlap. Because some users may want to compare diarization with and without overlap (but otherwise identical), we used the exact same utterances and alignments as above, with only one difference – in the overlap version we subtract 200 ms from each between-speaker interval. This shifts the mode of the ~Rayleigh distribution to 0 ms, with a range of -198 to 619 ms (mean 40 ms). This is a fairly realistic range of overlap for natural English conversations [37,16], and therefore barely noticeable to the human ear. Note, however, that real-world conversations also include another type of overlap, where one speaker makes a brief utterance or non-speech sound in the middle of the other speaker's turn (sometimes called "back-channel" speech). We have no statistics for such events, and it is not possible to imitate these easily with just the LibriSpeech data, so no such "back-channel" speech was included in the synthetic corpus.

Next, a 3-person synthetic dialog corpus was constructed by the same methods as above. However, we do not want all dialogs to have $\sim 33\%$ representation of each of the 3 speakers. Although we do not know of any published statistics, it is certainly not the case that all real-world 3-person dialogs have equal time allocated to the 3 speakers. Also, the 3 speakers should not alternate in a

Table 2. Synthetic 2-person corpus with overlap

	Dialogs	Utts (Turns)	Tokens	Hours
Train	292	28522	989715	96.58
Dev-clean	48	2673	53765	4.83
Dev-other	45	2822	50227	4.54
Test-clean	43	2605	52279	4.93
Test-other	45	2861	51305	4.69

simple sequence of 1, 2, 3, 1, 2, 3, etc. As a simple first solution, the sequence was assigned as follows: the first speaker is speaker 1 by definition, and then each subsequent speaker is chosen randomly from the other 2 speakers, until one speaker runs out of available utterances. In this manner, each dialog ends up with a unique sequence of speaker turns, and unique proportions of representation across the 3 speakers. Single speakers took a range of 17.7-44.4% of the dialog turns (mean 33.3%). This method does, however, lose some utterances in each dialog, so the total hours in the corpus is less than for the 2-speaker corpus (Table 3). Dialogs included between 17 and 366 utterances (median 118), and ranged in duration from 2.8-71.5 min (median 24.4 min). Across all 3-speaker dialogs, 22% were same-gender (m-m-m or f-f-f) and 78% were mixed-gender.

Table 3. Synthetic 3-person corpus without overlap

	Dialogs	Utts (Turns)	Tokens	Hours
Train	195	26694	928346	92.11
Dev-clean	32	2430	48899	4.53
Dev-other	30	2560	45664	4.26
Test-clean	29	2406	47639	4.61
Test-other	30	2684	48025	4.53

The inter-speaker intervals were again chosen randomly according to a Rayleigh distribution with mode of 200 ms (as above), and the actual samples ranged from 1 to 803 ms (mean 242 ms). To create the corresponding 3-person corpus with overlap (Table 4), the identical sequences and values were used, except with 200 ms subtracted from the inter-speaker intervals. This yielded intervals of -199 to 603 ms (mean 42 ms).

Table 4. Synthetic 3-person corpus with overlap

	Dialogs	Utts (Turns)	Tokens	Hours
Train	195	26694	928346	90.64
Dev-clean	32	2430	48899	4.40
Dev-other	30	2560	45664	4.12
Test-clean	29	2406	47639	4.47
Test-other	30	2684	48025	4.38

3 Discussion and Conclusion

A synthetic corpus of dialogs was made from the open-source LibriSpeech corpus and released for download:

https://github.com/EMRAI/emrai-synthetic-diarization-corpus.

The corpus includes timing information in several formats, and includes phoneme as well as speaker segmentations. Both 2-speaker and 3-speaker corpora, with and without overlap, are provided. In the future, we will likely add a 4-speaker corpus. Note that dialogs with different numbers of speakers can be combined by a user to obtain a data set where the number of speakers is not fixed.

As a synthetic corpus, there are several deviations from real-world data. First, there is very little background noise (but users could add their own for a better approximation to real conditions [21,46]). Second, conversational statistics were approximately mimicked, but cannot be considered perfectly realistic. Third, we included no intervals of truly multi-speaker speech, i.e., “back-channel” utterances by one speaker that occur fully within the turn of another speaker. Fourth, the LibriSpeech corpus itself consists of high-quality readings of audio books, which has certain advantages (such as high-quality phonetic alignments), but also makes the speech unrealistic to most real-world applications. Fifth, although our corpus is gender-balanced, we include no child or other special categories of speech. Finally, we only include 2-speaker and 3-speaker dialogs (and 4-speaker dialogs will be included in a future release).

Thus, we explicitly do NOT suggest that the synthetic corpus replaces the need for real-world data; applied workers must also obtain data for each particular application. Nonetheless, we believe that our general-purpose corpus serves as a useful starting point for diarization research, particularly in the early stages of system development, where a very challenging corpus peculiar to one recording situation is often less desirable. We advise the beginning researcher to attempt first the 2-speaker corpus without overlap, and then move on to consider overlap and more speakers, along with real-world data. It is, however, possible that training on this corpus can produce models that generalize to real-world situations (as in our companion paper).

References

1. Anguera Miró, X.: Robust speaker diarization for meetings. Ph.D. thesis, Univ. Politècnica de Catalunya (2006)
2. Anguera Miró, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O.: Speaker diarization: a review of recent research. IEEE Trans Audio Speech Lang Process **20**(2), 356–370 (2012)
3. Anguera Miró, X., Hernando Pericás, F.: Evolutive speaker segmentation using a repository system. In: Proc ICSLP. pp. 605–608. ISCA (2004)
4. Anguera Miró, X., Wooters, C., Peskin, B., Aguiló, M.: Robust speaker segmentation for meetings: the ICSI-SRI Spring 2005 diarization system. Lect Notes Comput Sci **3869**, 402–414 (2006)
5. Bozonnet, S., Vipperla, R., Evans, N.: Phone adaptive training for speaker diarization. In: Proc INTERSPEECH. pp. 494–497. ISCA (2012)
6. Burger, S., MacLaren, V., Yu, H.: The ISL meeting corpus: the impact of meeting type on speech style. In: Proc ICSLP. pp. 301–304. ISCA (2002)
7. Chen, I.F., Cheng, S.S., Wang, H.M.: Phonetic subspace mixture model for speaker diarization. In: Proc INTERSPEECH. pp. 2298–2301. ISCA (2010)
8. Delacourt, P., Kryze, D., Wellekens, C.: Speaker-based segmentation for audio data indexing. In: Proc ESCA Tutorial and Research Workshop. pp. 78–83. ISCA (1999)
9. Finley, G., Edwards, E., Robinson, A., Sadoughi, N., Fone, J., Miller, M., Suendermann-Oeft, D.: An automated medical scribe for documenting clinical encounters. In: Proc NAACL. ACL (2018)
10. Gangadharaiah, R., Narayanaswamy, B.: A novel method for two-speaker segmentation. In: Proc ICSLP. pp. 2337–2340. ISCA (2004)
11. Garofolo, J., Laprun, C., Michel, M., Stanford, V., Tabassi, E.: The NIST meeting room pilot corpus. In: Proc LREC. p. 4 p. ELRA (2004)
12. Gauvain, J.L., Adda, G., Lamel, L., Adda-Decker, M.: Transcribing broadcast news: the LIMSI Nov96 Hub4 system. In: Proc DARPA Speech Recognition Workshop. pp. 56–63. DARPA (1997)
13. Gish, H., Siu, M.H., Rohlicek, J.: Segregation of speakers for speech recognition and speaker identification. In: Proc ICASSP. vol. 2, pp. 873–876. IEEE (1991)
14. Godfrey, J., Holliman, E., McDaniel, J.: SWITCHBOARD: telephone speech corpus for research and development. In: Proc ICASSP. vol. 1, pp. 517–520. IEEE (1992)
15. Hain, T., Burget, L., Dines, J., McCowan, I., Garau, G., Karafiat, M., Lincoln, M., Moore, D., Wan, V., Ordelman, R., Renals, S.: The development of the AMI system for the transcription of speech in meetings. In: Proc Workshop MLMI. vol. LNCS 3869, pp. 344–356. Springer (2005)
16. Heldner, M., Edlund, J.: Pauses, gaps and overlaps in conversations. J Phon **38**(4), 555–568 (2010)
17. Hsieh, C.H., Wu, C.H., Shen, H.P.: Adaptive decision tree-based phone cluster models for speaker clustering. In: Proc INTERSPEECH. pp. 861–864. ISCA (2008)
18. Ikbal, S., Visweswariah, K.: Learning essential speaker sub-space using hetero-associative neural networks for speaker clustering. In: Proc INTERSPEECH. pp. 28–31. ISCA (2008)
19. Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, C.: The ICSI meeting corpus. In: Proc ICASSP. vol. 1, pp. 364–367. IEEE (2003)

20. Jothilakshmi, S., Ramalingam, V., Palanivel, S.: Speaker diarization using autoassociative neural networks. Eng Appl Artif Intell **22**(4-5), 667–675 (2009)
21. Kim, K., Kim, M.: Robust speaker recognition against background noise in an enhanced multi-condition domain. IEEE Trans Consum Electron **56**(3), 1684–1688 (2010)
22. Liu, C., Yan, Y.: Speaker change detection using minimum message length criterion. In: Proc ICSLP. pp. 514–517. ISCA (2000)
23. Meinedo, H., Neto, J.: A stream-based audio segmentation, classification and clustering pre-processing system for broadcast news using ANN models. In: Proc INTERSPEECH. pp. 237–240. ISCA (2005)
24. Metzger, Y.: Blind segmentation of a multi-speaker conversation using two different sets of features. In: Proc Odyssey Workshop. pp. 157–162. ISCA (2001)
25. Moattar, M., Homayounpour, M.: A review on speaker diarization systems and approaches. Speech Commun **54**(10), 1065–1103 (2012)
26. Mohammadi, S., Sameti, H., Langarani, M., Tavanaei, A.: KNNDIST: a nonparametric distance measure for speaker segmentation. In: Proc INTERSPEECH. pp. 2282–2285. ISCA (2012)
27. NIST: Spring 2006 (RT-06S) Rich Transcription Meeting Recognition Evaluation plan. Report RT-06S, National Institute of Standards and Technology (Spring 2006)
28. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: LibriSpeech: an ASR corpus based on public domain audio books. In: Proc ICASSP. pp. 5206–5210. IEEE (2015)
29. Povey, D., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlícek, P., Qian, Y., Schwarz, P., Silovsky, J.: The Kaldi speech recognition toolkit. In: Proc Workshop ASRU. p. 4 p. IEEE, Waikoloa Village, HI (2011)
30. Rohlicek, J., Ayuso, D., Bates, M., Bobrow, R., Boulanger, A., Gish, H., Jeanrenaud, P., Meteer, M., Siu, M.H.: Gisting conversational speech. In: Proc ICASSP. vol. 2, pp. 113–116. IEEE (1992)
31. Schindler, C., Draxler, C.: Using spectral moments as a speaker specific feature in nasals and fricatives. In: Proc INTERSPEECH. pp. 2793–2796. ISCA (2013)
32. Shoup, J.: Phonological aspects of speech recognition. In: Lea, W. (ed.) Trends in speech recognition, pp. 125–138. Prentice-Hall, Englewood Cliffs, NJ (1980)
33. Siegler, M., Jain, U., Raj, B., Stern, R.: Automatic segmentation, classification and clustering of broadcast news audio. In: Proc DARPA Speech Recognition Workshop. pp. 97–99. DARPA (1997)
34. Siu, M.H., Yu, G., Gish, H.: An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers. In: Proc ICASSP. vol. 2, pp. 189–192. IEEE (1992)
35. Soldi, G., Bozonnet, S., Alegre, F., Beaugeant, C., Evans, N.: Short-duration speaker modelling with phone adaptive training. In: Proc Odyssey Workshop. pp. 208–215. ISCA (2014)
36. Sönmez, M., Heck, L., Weintraub, M.: Speaker tracking and detection with multiple speakers. In: Proc EUROSPEECH. pp. 2219–2222. ISCA (1999)
37. Stivers, T., Enfield, N., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J., Yoon, K.E., Levinson, S.: Universals and cultural variation in turn-taking in conversation. Proc Natl Acad Sci U S A **106**(26), 10587–10592 (2009)
38. Sugiyama, M., Murakami, J., Watanabe, H.: Speech segmentation and clustering based on speaker features. In: Proc ICASSP. vol. 2, pp. 395–398. IEEE (1993)
39. Takagi, K., Itahashi, S.: Segmentation of spoken dialogue by interjections, disfluent utterances and pauses. In: Proc ICSLP. pp. 697–700. ISCA (1996)

40. Valente, F., Wellekens, C.: Scoring unknown speaker clustering: VB vs. BIC. In: Proc ICSLP. pp. 593–596. ISCA (2004)
41. Viñals, I., Villalba, J., Ortega, A., Miguel, A., Lleida, E.: Bottleneck based front-end for diarization systems. Lect Notes Comput Sci **10077**, 276–286 (2016)
42. Wang, G., Wu, X., Zheng, T.: Using phoneme recognition and text-dependent speaker verification to improve speaker segmentation for Chinese speech. In: Proc INTERSPEECH. pp. 1457–1460. ISCA (2010)
43. Wilcox, L., Chen, F., Kimber, D., Balasubramanian, V.: Segmentation of speech using speaker identification. In: Proc ICASSP. vol. 1, pp. 161–164. IEEE (1994)
44. Yella, S., Motlícek, P., Bourlard, H.: Phoneme background model for information bottleneck based speaker diarization. In: Proc INTERSPEECH. pp. 597–601. ISCA (2014)
45. Yella, S., Stolcke, A., Slaney, M.: Artificial neural network features for speaker diarization. In: Proc SLT Workshop. pp. 402–406. IEEE (2014)
46. Zâo, L., Coelho, R.: Colored noise based multicondition training technique for robust speaker identification. IEEE Signal Process Lett **18**(11), 675–678 (2011)
47. Zibert, J., Mihelic, F.: Prosodic and phonetic features for speaker clustering in speaker diarization systems. In: Proc INTERSPEECH. pp. 1033–1036. ISCA (2011)