The EURECOM submission to the first DIHARD Challenge

Jose Patino, Héctor Delgado and Nicholas Evans

Department of Digital Security, EURECOM, Sophia Antipolis, France

{patino,delgado,evans}@eurecom.fr

Abstract

The first DIHARD challenge aims to promote speaker diarization research and to foster progress in domain robustness. This paper reports EURECOM's submission to the DIHARD challenge. It is based upon a low-resource, domain-robust binary key approach to speaker modelling. New contributions include the use of an infinite impulse response - constant Q Mel-frequency cepstral coefficient (ICMC) front-end, a clustering selection / stopping criterion algorithm based on spectral clustering and a mechanism to detect single-speaker trials. Experimental results obtained using the standard DIHARD database show that the contributions reported in this paper deliver relative improvements of 39% in terms of the diarization error rate over the baseline algorithm. An absolute DER of 29% on the evaluation set compares favourably with those of competing systems, especially given that the binary key system is highly efficient, running 63 times faster than real-time.

Index Terms: Speaker diarization, binary key, spectral clustering, zero-resource diarization

1. Introduction

While speaker diarization research attracted significant interest in the past, the field has somewhat stagnated in recent times. This is perhaps due to the lack of significant datasets; those used in the NIST Rich Transcription evaluations [1] contained only a small number of recordings which makes the comparison of different technologies rather difficult. Even more recent databases such as those used for the ETAPE [2] and REPERE [3] evaluations are modest in size. All of these datasets are furthermore narrow in terms of their application scenario, e.g. broadcast news, meetings or televised chat shows. As a result, each database and evaluation has a somewhat limited audience.

The DIHARD initiative [4] was born to re-energise the research effort. The availability of a larger, standard dataset supporting a broader range of application scenarios, e.g. including medical interviews, conversations involving children, even monologues, stands to rejuvenate research interest and especially to foster progress in domain-robust speaker diarization; the DIHARD dataset contains no training data and represents the broadest domain variation captured in a single speaker diarization dataset to date.

There are two distinct approaches to address such a challenge. The first entails the optimisation of systems using a large quantity of training data that spans adequately the domain variation captured in the DIHARD data. The second is an inherently domain-neutral approach that requires no background training data, or rather acquires background data from acoustic streams at runtime. A hybrid approach might aim to exploit the benefit of background training data, but with the facility to adapt to a specific domain at runtime.

Given our interest in low-resource and computationally efficient, practicable speaker diarization technology, our efforts to address the first DIHARD challenge have explored the second approach. Past work has shown the merit of a so-called binary key approach to speaker diarization [5] that does not require any background training data. It has been applied successfully to the diarization of variable domain data [6] and operates substantially faster than realtime.

Since it does not require background training data, it is ideally suited to domain-robust diarization. However, while its principal merit relates to computational efficiency, rather than raw performance, it is not necessarily expected to be competitive with the best-performing submissions to the first DIHARD challenge. Results show nonetheless that, with the introduction of three modifications, it remains surprisingly competitive.

Modifications involve new front-end processing, a new clustering selection / stopping criterion and a mechanism to detect single-speaker trials. The front-end uses infinite impulse response - constant Q Mel-frequency cepstral coefficients (ICMCs) developed originally for automatic speaker verification [7]. Cluster selection is based upon spectral clustering, shown by other authors to improve diarization performance by estimating more reliably the number of speakers. Single-speaker detection is found to be beneficial in reducing errors attributed to the under-clustering of single-speaker trials.

The remainder of this paper is organised as follows. The original binary key approach to speaker diarization is described in Section 2. Enhancements to the baseline systems are described in Section 3. Experiments and results are described in Sections 4 and 5. Conclusions are presented in Section 6.

2. Baseline system

EURECOM's submission to the DIHARD challenge is based on a binary key (BK) modelling technique. Applied originally to speaker recognition [8, 9], BK modelling has since been applied to a variety of different tasks such as speech activity detection [10], emotion recognition [11] and to speaker diarization and related tasks [5, 10, 12–15].

2.1. Binary key modelling

The merit of BK modelling for speaker diarization lies in computational efficiency. Whereas many competing approaches such as conventional agglomerative hierarchical clustering (AHC) techniques are often computationally impracticable, BK-based approaches run substantially faster than realtime [5, 10, 15]. BK-based approaches to speaker diarization are also flexible in their use of background training data and can operate entirely *without* training data [5, 6], learning necessary background information at runtime.

By making no assumptions about the domain, this particular quality of BK-based approaches to diarization make it particularly well suited to domain-robust diarization. This is especially so given that the diarization task is often characterised by substantial variability such as acoustic content, number of speakers

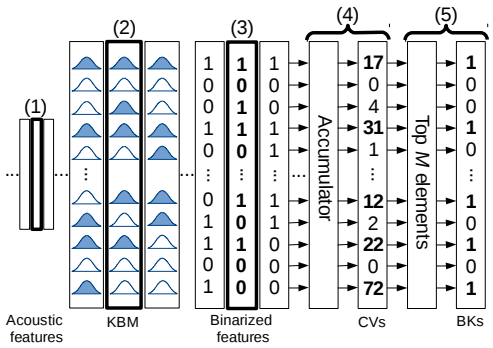


Figure 1: *An illustration of the BK extraction procedure based upon the comparison of acoustic features and the KBM.*

and differences in speaker floor time, all examples of variation which characterise the DIHARD dataset.

2.2. The binary key background model

BK-modelling relies on a so-called binary key background model (KBM). Its purpose is similar in nature to that of a traditional universal background model (UBM) [16]. The KBM is learned using traditional acoustic features, i.e. MFCCs in the original work [5]. The feature stream is then segmented into 2-second windows and a single, multi-variate Gaussian distribution is then fitted to the set of frames in each window. This generates an over-sampled representation of the acoustic space in the form of a pool of multi-variate Gaussian components. In order to remove redundancy, the pool is decimated according to the recursive procedure described in [5]. This procedure is performed by measuring iteratively the cosine similarity between each already selected component and the remaining components in the pool. The most dissimilar Gaussian among those in the pool is extracted and added to the KBM. The procedure is performed iteratively to give a KBM of N components.

2.3. Cumulative vector / binary key extraction

Binarised features are obtained from the comparison of acoustic features with the KBM. The process is illustrated in Fig. 1. A sequence of n_f acoustic features is transformed into a binary key (BK) whose dimension N is dictated by the number of components in the KBM. For each acoustic feature vector (labelled 1 to the left of Fig. 1), the likelihood given each of the N KBM components is computed and stored in a vector which is sorted by Gaussian index. The top N_G Gaussians defined as those with the N_G highest likelihoods (2 - illustrated in solid blue) are then selected and used to create binarised versions of the acoustic features (3).

This process is repeated for each frame of acoustic features thereby resulting in a binary matrix of dimension $n_f \times N$, each column of which has N_G values equal to binary 1. A row-wise addition of this matrix is then used to determine a single cumulative vector (CV) which reflects the number of times each Gaussian in the KBM was selected as a top-Gaussian (4). The final BK is obtained from the M positions with highest values in the CV (5). Corresponding elements in the BK are set to binary 1 whereas others are set to 0. Both the CV and the BK provide a sparse, fixed length representation of a speech segment based on similarity to the acoustic space modelled by the KBM. Diarization can be performed using CVs or BKs. The work in [5] shows that CVs perform best. Accordingly, all work presented in this paper was performed using CV features.

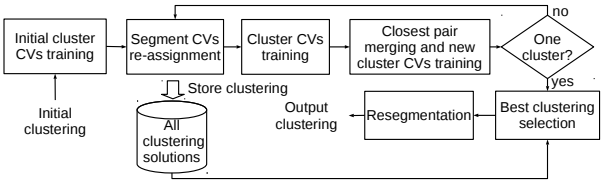


Figure 2: *The baseline, bottom-up agglomerative hierarchical clustering process.*

2.4. Diarization system

The baseline speaker diarization system is that described in [6] and illustrated in Fig. 2. It is based upon a bottom-up, agglomerative hierarchical clustering (AHC) algorithm with cluster model re-training and segment reassignment. According to cumulative vector extraction procedure described in Section 2.3, test data is first converted into a sequence of CVs each representing contiguous intervals of 3s, overlapped by 2s. The AHC algorithm is then initialised with the definition of $N_{init} = 25$ contiguous, equally sized segments. Upon initialisation, the acoustic features in each segment are used to define a single *cluster* CV following the same CV extraction algorithm.

The iterative AHC algorithm is then applied to the set of segment and cluster CVs, with the number of clusters being reduced by one upon each iteration. Segment CVs are compared one-by-one to cluster CVs using the cosine similarity and are re-assigned to the *closest* cluster. Cluster CVs are then re-estimated from cluster contents.

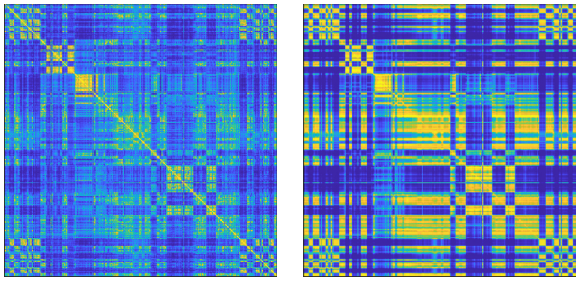
Using a cluster CV cosine similarity matrix, all segments assigned to the two closest clusters are then reassigned to a new, merged cluster CV such that the number of clusters decreases by one. The re-segmentation and re-estimation process is performed iteratively, each time followed by a step of cluster merging until there remains only a single cluster. A cluster selection algorithm is then applied to determine the *best* number of clusters, i.e. the number of speakers. As described in [15], selection is performed using an elbow criterion which is applied to the curve of the within-class sum-of-squares (WCSS) of all clustering solutions, with the goal of finding a trade-off between the number of clusters and cluster dispersion. A final maximum likelihood re-segmentation using 128-component Gaussian mixture models (GMM) is performed at the acoustic feature level in order to refine the segmentation of the selected clustering.

3. Baseline enhancements

This paper reports three enhancements to the baseline system that were found to improve performance in the face of domain variability. They involve the use of different acoustic features, a different cluster selection / stopping criterion and an approach to detect single-speaker documents.

3.1. Acoustic features

In place of baseline MFCC acoustic features, EURECOM's submission to the first DIHARD challenge uses infinite impulse response, constant Q transform Mel-frequency cepstral coefficients (ICMC). ICMC features were explored originally in the context of speaker recognition and utterance verification [10]. ICMC extraction performs spectro-temporal decomposition with the constant Q transform [17], a perceptually motivated alternative to the short-time Fourier transform (STFT). Whereas the spectro-temporal resolution of the STFT is con-



(a) Original (b) Refined

Figure 3: *Affinity matrix of cosine similarities between CVs for the file D_0028.wav before (a) and after (b) the refinement process.*

stant, that of the CQT exhibits a constant Q factor. The Q factor is a measure of the filter selectivity and is defined as the ratio between the centre frequency and the bandwidth. A constant Q factor gives greater spectral resolution at lower frequencies, and greater temporal resolution at higher frequencies. The human perception system approximates a constant Q factor between 500Hz and 20kHz. This is the principal motivation behind the use of the CQT for speech and audio analysis [18–21].

Unfortunately, compared to the STFT, the CQT is computationally expensive. In order to mitigate additional computation, all work reported in this paper was performed using the infinite impulse response - constant Q transform (IIR-CQT) algorithm proposed in [22]. The IIR-CQT, a compromise between computational cost and design flexibility, gives a constant Q spectro-temporal decomposition through IIR linear time variant filtering of the STFT. Full details can be found in [22]. ICMC features are then obtained from the IIR-CQT with traditional Mel-cepstral analysis[1].

3.2. Clustering

Motivated by other reports of its successful application to speaker diarization [23–25], we explored the use of spectral clustering [26] with binary key modelling. Spectral clustering overcomes the drawbacks of more conventional approaches to clustering that are linked to the use of parametric density estimators and optimisation to local minima. The general idea is to perform clustering using the eigenvectors corresponding to the top eigenvalues estimated from an *affinity matrix* derived from the similarities between data points being clustered, i.e. the CVs.

All work reported in this paper was preformed using the spectral clustering algorithm presented in [25]. This work proposes a number of refinements to the affinity matrix that are applied prior to eigenvalue decomposition and that give improved speaker diarization performance. They are based on the temporal locality of speech data. Contiguous speech segments uttered by the same speaker should have similar CVs and hence similar values in the affinity matrix. Given a test audio file, represented by a sequence of M segment CVs, the M-by-M affinity matrix is determined using the cosine similarity and then treated by a series of operations including Gaussian blurring with standard deviation σ, row-wise thresholding of similarities below the p-percentile, symmetrisation, diffusion and row-wise Max normalisation. Full details of each are given in [25].

Fig. 3 shows an example of an affinity matrix computed for DIHARD development set file *DH_0028.wav* (a) before and (b) after refinement which smooths and denoises the data in the similarity space. Even if the original affinity matrix already highlights patterns corresponding to different speakers and turns, these are more uniform and sharper after refinement. These improvements are crucial to subsequent eigenvalue decomposition.

Eigenvalue decomposition is then performed and the eigenvalues are sorted in descending order: $\lambda_1 > \lambda_2 > ... > \lambda_n$. The number of clusters $\widetilde{k}$ is then selected according to the value k which maximises the eigengap defined as:

$$\widetilde{k} = \arg \max_{1 \leq k \leq n} \frac{\lambda_k}{\lambda_{k+1}} \quad (1)$$

The M-by-$\widetilde{k}$ matrix of eigenvectors corresponding to the $\widetilde{k}$ largest eigenvalues is then used to obtain a $\widetilde{k}$-dimensional representation of the M input CVs. These lower dimensional representations are then clustered with a k-means algorithm using the squared Euclidean distance. Since the algorithm estimates both the number of clusters and the clustering, it can also be used only as a stopping criterion for other clustering algorithms.

3.3. Single-speaker detection

Finally, having found that diarization errors in single-speaker documents produce high error rates, we designed a specific mechanism for single-speaker detection. Since the spectral clustering algorithm described above results too often in the estimation of a single speaker, it is configured to force the return of two or more clusters. Single-speaker detection is then performed pre-clustering according to the thresholding of the eigengap between the two largest eigenvalues. In the case that $\lambda_1 - \lambda_2$ exceeds a threshold θ, then the number of clusters is forced to 1.

4. Experimental setup

All experimental work reported in this paper was performed with the standard DIHARD database [4, 27, 28]. The development set contains 164 audio documents from 9 different domains. All results correspond to the use of ground-truth speech activity detection annotations, i.e. track 1 of the DIHARD challenge.

Baseline acoustic features are MFCCs comprising 19 static coefficients computed from windows of 25ms with 10ms overlap and with a filterbank of 20 channels. ICMC features use longer windows of 128ms, also with 10ms overlap. The KBM is determined from a pool of Gaussians, each estimated using windows of between 0.5 to 2 seconds duration set dynamically so as to ensure a minimum of 1024 components. The size of the KBM after Gaussian selection is set to an empirically optimised percentage of the number in the original pool, details of which are presented later. Segment CVs are estimated using 3s windows with 2s overlap. The top number of Gaussians per frame is set to $N_G = 5$.

AHC clustering is initialised with $N_{init} = 25$ clusters. Based on the distribution of the number of speakers per document on the development set, the maximum number of output clusters is set to 10. For spectral clustering, only eigenvalues larger than a threshold $\delta = 2.1$ are used to compute eigengaps to decide the number of clusters. This is done after observing that very low eigenvalues may produce anomalously large eigengaps, resulting in excessive clusters. The single-speaker detection threshold is set to $\theta = 410$. System performance is

[1] A Matlab implementation is available at `http://audio.eurecom.fr/content/software`

Table 1: *Speaker diarization performance in terms of diarization error rate (DER, %) of the baseline system and after incorporating the proposed enhancements, on the development and evaluation sets. DER is also broken-down by domain for the development set (D1: SEEDLINGS, D2: SCOTUS, D3: DCIEM, D4: ADOS, D5: YP, D6: SLX, D7: RT04S, D8: LIBRIVOX, D9: VAST).*

	Development										Eval.
Systems	D1	D2	D3	D4	D5	D6	D7	D8	D9	ALL	ALL
1. MFCC / AHC / elbow (baseline)	59.64	**8.36**	44.35	46.38	28.34	46.97	46.99	66.69	56.75	44.47	48.31
2. ICMC / AHC / elbow	44.85	9.37	46.05	46.58	24.39	49.49	46.02	66.97	59.63	44.85	48.70
3. ICMC / SC	48.68	17.31	17.85	31.02	**11.36**	**23.65**	43.04	27.31	45.16	30.13	34.29
4. ICMC / AHC / SC$_{\#spk}$	**43.78**	14.19	**9.70**	**27.48**	12.71	23.99	**42.24**	11.22	38.33	25.77	30.44
5. ICMC / AHC / SC$_{\#spk}$ / 1-spk	**43.78**	14.19	11.02	**27.48**	12.71	23.99	43.55	**5.36**	**38.24**	**25.56**	**29.33**

assessed using the standard diarization error rate (DER) with no forgiveness collar. Intervals containing overlapping speech regions are also scored.

5. Results

Results presented in Table 1 show diarization performance measured in terms of the DER for both development and evaluation sets. Results are shown for the baseline system (line 1) and for the same system with the proposed enhancements (lines 2-5). The baseline system uses MFCC features, standard AHC and elbow cluster selection. It achieves a DER of almost 50% for the evaluation set.

System 2 is identical except for the use of ICMC features. While these results show that ICMC feature give worse performance than MFCC features, results illustrated Fig. 4 indicate otherwise. Fig. 4 plots DER results against the KBM size for MFCC and ICMC features for oracle cluster selection. For smaller KBM sizes, ICMC produces substantially lower DERs than MFCC features. Thus, assuming more reliable cluster selection, then ICMC features will give lower DERs. This hypothesis is confirmed by the results of subsequent experiments.

System 3 uses spectral clustering (SC) in place of AHC and elbow cluster selection. The use of SC leads to a 29% relative reduction in DER over the baseline for the evaluation set. System 4 combines AHC clustering with the SC selection algorithm (SC$_{\#spk}$) Performance improves again, this time giving a relative improvement of 37% over the baseline system. System 5 is identical to system 4 except for the application of the single-speaker detection mechanism reported in Section 3.3. This system gives the best performance, with an absolute DER of 29% corresponding to a relative improvement of almost 40% over the baseline.

Table 1 also shows granular results for each of the 9 domains D1-D9 (consult [4] for details) contained in the development set. Diarization performance improves for most domains with the application of the proposed system enhancements. The exception is D2, for which the baseline system performs the best. This is attributed to the tendency of the spectral clustering selection algorithm to underestimate the number of clusters. Of particular note are improvements for D8. Documents corresponding to this domain contain only a single-speaker. Here, the single-speaker detection mechanism is especially effective in reducing the error rate.

While overall performance is lower than the best submissions to the DIHARD challenge, results are surprisingly competitive. The system proposed in this paper offers an appealing compromise between performance and efficiency. System 5 processes the entire development set in 835s running on an Intel Core i5-3470 3.20GHz CPU with 16GB of RAM. This corresponds to real-time factor of 0.016 (63 times faster than real-time). Few other systems are so computationally efficient.

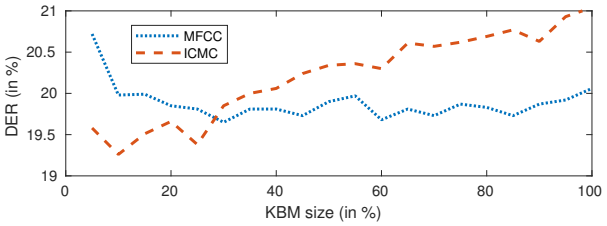


Figure 4: *Comparison of MFCC and ICMC features for different KBM sizes when using an oracle selection of clustering solutions (those which minimise DER).*

6. Conclusions

This paper reports EURECOM's submission to the first DIHARD challenge in domain-robust speaker diarization. While the baseline system is shown to perform poorly, the three enhancements reported in this paper lead to substantial improvements over the baseline system. Enhancements include features extracted using a perceptually motivated, variable spectro-temporal decomposition, a robust approach to cluster selection based upon spectral clustering and a mechanism designed to detect single-speaker segments. When combined, these enhancements bring a relative reduction in the diarization error rate of almost 40% over the baseline system. Performance, although lower than that of top-ranked systems, still compares favourably. This is especially so given that the proposed system requires no background data and is highly efficient, with execution times in order of 63 times faster than real time when run on a consumer-grade desktop computer.

With respect to the goal of domain-robustness, the proposed system based on binary-key modelling is a ready-to-run or off-the-shelf solution to speaker diarization. The estimates of speaker diarization performance reported in this paper are likely to be reasonably reliable estimates of performance if the same system were to be tested using data collected in other domains; the system is not dependent on optimisation using domain-specific background data and is instead tuned at runtime. This is seen as a significant advantage over competing systems. This quality should be of appeal to practical applications of speaker diarization technology which is, after all, often an enabling technology rather than the final application.

7. References

[1] "The NIST Rich Transcript Evaluation 2009," 2009, https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation.

[2] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert, "The ETAPE corpus for the evaluation of speech-based TV content processing in the French language," in *LREC-Eighth international conference on Language Resources and Evaluation*, 2012, p. na.

[3] O. Galibert and J. Kahn, "The first official REPERE evaluation," in *First Workshop on Speech, Language and Audio in Multimedia*, 2013.

[4] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First DIHARD challenge evaluation plan," 2018, https://zenodo.org/record/1199638.

[5] H. Delgado, X. Anguera, C. Fredouille, and J. Serrano, "Fast single-and cross-show speaker diarization using binary key speaker modeling," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 12, pp. 2286–2297, 2015.

[6] J. Patino, H. Delgado, N. Evans, and X. Anguera, "EURECOM submission to the Albayzin 2016 Speaker Diarization Evaluation," in *Proc. IberSPEECH*, 2016.

[7] H. Delgado, M. Todisco, M. Sahidullah, A. K. Sarkar, N. Evans, T. Kinnunen, and Z.-H. Tan, "Further optimisations of constant Q cepstral processing for integrated utterance and text-dependent speaker verification," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 179–185.

[8] X. Anguera and J.-F. Bonastre, "A novel speaker binary key derived from anchor models," in *Proc. INTERSPEECH*, 2010, pp. 2118–2121.

[9] J.-F. Bonastre, X. A. Miró, G. H. Sierra, and P.-M. Bousquet, "Speaker Modeling Using Local Binary Decisions," in *Proc. INTERSPEECH*, 2011, pp. 13–16.

[10] H. Delgado, C. Fredouille, and J. Serrano, "Towards a complete binary key system for the speaker diarization task," in *Proc. INTERSPEECH*, 2014, pp. 572–576.

[11] J. Luque and X. Anguera, "On the modeling of natural vocal emotion expressions through binary key," in *Proc. EUSIPCO*, 2014, pp. 1562–1566.

[12] X. Anguera and J.-F. Bonastre, "Fast speaker diarization based on binary keys," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4428–4431.

[13] H. Delgado, X. Anguera, C. Fredouille, and J. Serrano, "Global Speaker Clustering towards Optimal Stopping Criterion in Binary Key Speaker Diarization," in *Proc. IberSPEECH 2014*, 2014, pp. 59–68.

[14] H. Delgado, X. Anguera, C. Fredouille, and J. Serrano, "Improved binary key speaker diarization system," in *Proc. EUSIPCO*, 2015, pp. 2087–2091.

[15] H. Delgado, X. Anguera, C. Fredouille, and J. Serrano, "Novel Clustering Selection Criterion for Fast Binary Key Speaker Diarization," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 3091–3095.

[16] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[17] J. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, January 1991.

[18] G. Costantini, R. Perfetti, and M. Todisco, "Event Based Transcription System for Polyphonic Piano Music," *Signal Processing*, vol. 89, no. 9, pp. 1798–1811, Sep. 2009.

[19] R. Jaiswal, D. Fitzgerald, E. Coyle, and S. Rickard, "Towards shifted NMF for improved monaural separation," in *24th IET Irish Signals and Systems Conference (ISSC 2013)*, June 2013, pp. 1–7.

[20] C. Schorkhuber, A. Klapuri, and A. Sontacch, "Audio pitch shifting using the constant-Q transform," *Journal of the Audio Engineering Society*, vol. 61, no. 7/8, pp. 425–434, July/August 2013.

[21] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516 – 535, 2017.

[22] P. Cancela, M. Rocamora, and E. López, "An efficient multi-resolution spectral transform for music analysis," in *Proc. ISMIR*, 2009, pp. 309–314.

[23] S. Shum, N. Dehak, and J. Glass, "On the use of spectral and iterative methods for speaker diarization," in *Proc. INTERSPEECH*, 2012.

[24] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised Methods for Speaker Diarization: An Integrated and Iterative Approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, Oct 2013.

[25] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with LSTM," in *Proc. ICASSP*, Calgary, Canada, 2018.

[26] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, 2002, pp. 849–856.

[27] E. Bergelson, "Bergelson Seedlings HomeBank Corpus," 2016, doi: 10.21415/T5PK6D.

[28] N. Ryant *et al.*, "DIHARD Corpus," 2018, linguistic Data Consortium.