# Pitch-synchronous Discrete Cosine Transform Features for Speaker Identification and Verification

Amit Meghanani[a] and A. G. Ramakrishnan[b]
*Indian Institute of Science, Bangalore, India*

Keywords: Pitch-synchronous, DCT, MFCC, Speaker Identification, Speaker Verification.

Abstract: We propose a feature called pitch-synchronous discrete cosine transform (PS-DCT), derived from the voiced part of the speech for speaker identification (SID) and verification (SV) tasks. PS-DCT features are derived from the 'time-domain, quasi-stationary waveform shape' of the voiced sounds. We test our PS-DCT feature on TIMIT, Mandarin and YOHO datasets. On TIMIT with 168 and Mandarin with 855 speakers, we obtain the SID accuracies of 99.4% and 96.1%, respectively, using a Gaussian mixture model-based classifier. In the i-vector-based SV framework, fusing the 'PS-DCT based system' with the 'MFCC-based system' at the score level reduces the equal error rate (EER) for both YOHO and Mandarin datasets. In the case of limited test data and session variabilities, we obtain a significant reduction in EER, up to 5.8% (for test data of duration $< 3$ sec).

## 1 INTRODUCTION

Traditional signal processing based features used in the literature for speaker identification (SID) and verification (SV) studies include Mel frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980), linear prediction residual (Prasanna et al., 2006), voice source cepstral coefficients (Gudnason and Brookes, 2008), deterministic plus stochastic model (Drugman and Dutoit, 2012) and MFCC with phase (Nakagawa et al., 2012), etc. Except the MFCC, all the above mentioned features are based on voice source. MFCC are motivated by human auditory perceptual mechanism and it primarily captures the vocal-tract information.

In contrast, we propose a feature, which exploits the time-domain, quasi-periodic waveform shape of the voiced phones. Since the voiced regions are relatively stationary, each period of any voiced phone must contain information from both the vocal folds and the vocal tract of the speaker, in addition to the phonemic information. This point of view motivated us to derive another traditional signal processing based feature, which is proposed in this work. To capture the time-domain waveform shape of the voiced phones, we employ discrete cosine transform (DCT), which is a reversible, linear transformation, with frequency resolution twice that of discrete Fourier transform. Since DCT has energy compaction property, it can capture the waveshape of any voiced phone for any pitch cycle in a limited number of coefficients.

The pitch-synchronous DCT analysis is obviously valid only for the voiced regions of speech. Pitch-synchronous analysis is important as DCT is shift-variant. Therefore, we avoid unvoiced segments of the speech. We consider each pitch cycle from the voiced segments as an analysis frame and perform DCT of a fixed length, after zero-padding the analysis frame. Hence, the DCT of different cycles for the same phone for a particular speaker are expected to show similar trends. Pitch synchronous analysis of voiced sounds was reported in (Mathews et al., 1961) to show that vowel sounds can be represented by a sequence of poles arising from the vocal tract and a sequence of zeros characterizing the glottal excitation. Similarly, DCTILPR (Abhiram et al., 2015), the pitch synchronous DCT of the integrated linear prediction residual (ILPR) was proposed for the task of speaker identification. In our work, DCT is applied on the speech signal itself as opposed to the case of DCTILPR, where DCT is applied on ILPR, which contains only the voice source information. Also, we use uniform length DCT basis as opposed to the above work (Abhiram et al., 2015), where variable length DCT basis is used.

[a] https://orcid.org/0000-0002-0811-274X
[b] https://orcid.org/0000-0002-3646-1955



Meghanani, A. and Ramakrishnan, A.
Pitch-synchronous Discrete Cosine Transform Features for Speaker Identification and Verification.
DOI: 10.5220/0008911503950401
In *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2020)*, pages 395-401
ISBN: 978-989-758-397-1; ISSN: 2184-4313

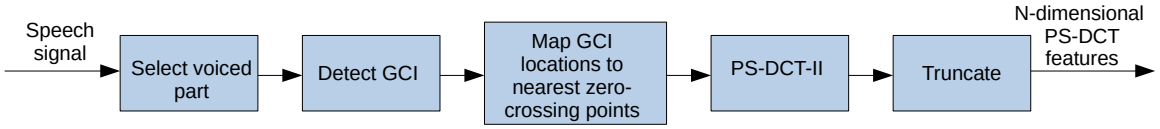


Figure 1: Extraction of the proposed PS-DCT features from the speech signal.

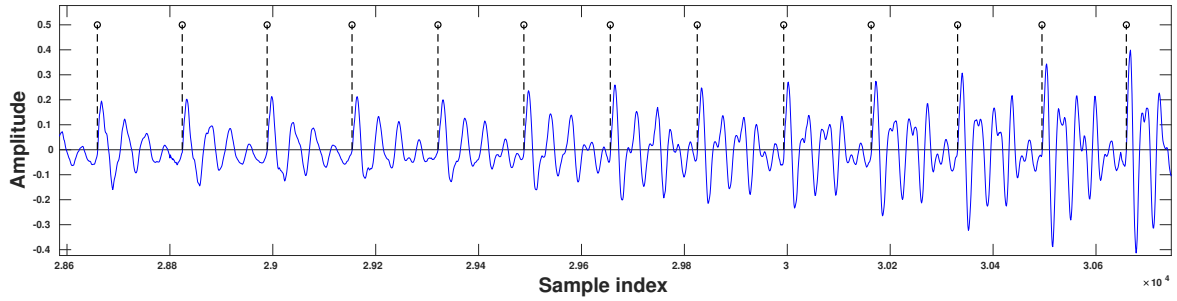


Figure 2: GCI locations of a voiced region of an utterance from the TIMIT dataset, after mapping them to the nearest zero-crossings (dotted vertical lines).

The major contributions of this paper are:

- Proposing a novel feature named PS-DCT, derived using traditional signal processing ideas, which can be used for speaker identification and verification.
- Showing that fusing the 'PS-DCT based SV system' with the 'MFCC-based SV system' improves the performance of speaker verification system for limited test data (see Table 4).

In a nutshell, we propose a feature named PS-DCT, which when combined with the conventional MFCC-based system, boosts the performance of the SV/SID system. PS-DCT possesses both voice source and vocal tract information, thus providing supplementary information to the MFCC-based SV/SID. This is crucial for short-utterances, since the performance drops significantly with less test data, even when the training data is sufficient (Das et al., 2014). By combining the systems based on MFCC and PS-DCT at the score level, good SV performance is obtained.

## 2 EXTRACTION OF THE PROPOSED PS-DCT FEATURES

Figure 1 shows the method of extraction of the PS-DCT features as a block diagram. Since we are extracting the features from the voiced part of the speech, we need to separate voiced and unvoiced speech. For this purpose, we have used normalized autocorrelation at unit sample delay. By using this method with empirically obtained threshold values, we take only voiced part of the speech and discard all the unvoiced/silence part of the speech. The algorithm proposed in (Prathosh et al., 2013) is used on these voiced regions for getting the glottal closure instants (GCI). The obtained GCI locations are then mapped to the nearest zero-crossings, which obviates the scenario of abrupt onset and ending of the signal in the analysis frames. In Fig. 2, the dotted vertical lines show the locations of these shifted GCIs. Our analysis frames correspond to the intervals between these zero-crossings.

The next block in Fig. 1 is PS-DCT of the analysis intervals. Let M be the number of samples in a pitch period. In simple pitch-synchronous analysis, the value of this analysis interval (M) changes with the pitch period, which can be different even for consecutive pitch cycles. If M-point DCT is used, then the DCT basis will be different for different analysis intervals, since M itself varies. Then, the k-th DCT coefficients of two different speakers (or even for the same speaker) do not represent the same frequency, but the same harmonic of their instantaneous fundamental frequency. The basis needs to be fixed for all the speakers, so that it can serve as a reference and we can obtain the relative distribution of the coefficients over the same basis to capture discriminative, speaker-specific information. In order to make the length of the analysis frame same (say, $L$) for all the pitch cycles for all the speakers, we choose the maximum number of samples that a pitch cycle can have as $L$. Thus $L$, the length of the basis of DCT, is determined as *(sampling rate)/(minimum pitch under consideration)*. The value of $L$ turns out to be 230 or 115, depending upon whether the sampling rate is 16 or 8 kHz, with the minimum pitch value assumed to

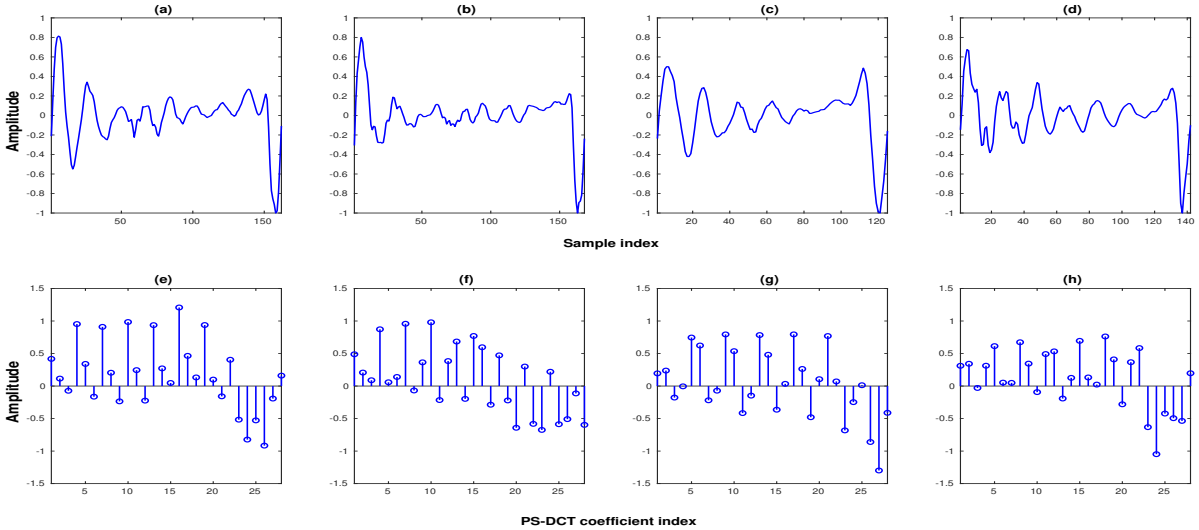


Figure 3: (a), (b), (c), (d): The time-domain waveform shape of one pitch period of the vowel /ah/ for four different speakers taken from the vowel database (https://homepages.wmich.edu/~hillenbr/voweldata.html). (e), (f), (g), (h) : 28-dimensional PS-DCT features derived from the waveforms shown in Figs. a, b, c, d, respectively.

Table 1: Choice of the number (**N**) of DCT coefficients retained vs SID accuracy (in %) on TIMIT database for 168 speakers and Chinese Mandarin corpus for 855 speakers.

| **No. of Coefficients** | **N** | 10 | 12 | 14 | 16 | 18 | 20 | 22 | 24 | 26 | **28** | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SID accuracy (in %)** | **TIMIT** | 87.5 | 90.4 | 93.4 | 94.6 | 95.8 | 95.8 | 96.4 | 96.4 | 97 | **99.4** | 97.6 |
| | **Mandarin** | 85.3 | 89.9 | 91.3 | 92.9 | 94.2 | 94 | 94.8 | 95.3 | 95.9 | **96.1** | 96.1 |

be 70 Hz. To the signal corresponding to each pitch period, we append appropriate number of zeros to get $L$ samples and take $L$-point DCT.

Before applying DCT, we normalize the analysis frame by its maximum absolute amplitude.

Then, we truncate the PS-DCT, ensuring that the retained $N$-coefficients have the necessary speaker-specific information. The value of $N$ is chosen experimentally, based on the best SID performance, to be 28 or 56, depending upon the sampling rate. The first coefficient, being the mean value of the waveform, is excluded.

For TIMIT and Mandarin corpora, we observe from Table 1 that the best value of N is 28, determined empirically. Similarly, N is 56 for YOHO database.

## 2.1 PS-DCT Captures Speaker-specific Information

The shapes of the time-domain quasi-periodic waveforms of a particular voiced phone are distinct for different speakers, though they look grossly similar as shown in Fig. 3 (a, b, c, d). Since PS-DCT captures the temporal waveform shape of a period, even for the same voiced phone (eg. /ah/), it is distinct for different speakers as shown in Fig. 3 (e, f, g, h). In other words, for a particular voiced phone, the PS-DCT gives the distinct distributions of the coefficients over the same fixed basis for different speakers, thus capturing discriminative speaker-specific information.

# 3 EXPERIMENTAL DETAILS OF THE STUDY

## 3.1 Speaker Identification (SID) Studies

Experiments are conducted on TIMIT (Garofolo et al., 1993), YOHO (Campbell and Higgins, 1994) and Chinese Mandarin (https://openslr.org/38/) databases for text-independent SID. TIMIT consists of data from 630 speakers, sampled at 16 kHz. For our study, we use data from randomly selected 168 speakers. Of the ten utterances available for each speaker, feature vectors obtained from 8 sentences are used for training and 2 sentences for testing. Chinese Mandarin corpus has a total of 102600 utterances from 855 speakers. Each speaker has 120 utterances, sampled at the rate of 16 kHz. For all the speakers, 8 utterances

are used for training and features obtained from 2 utterances, for testing. The YOHO database with four different recording sessions is used to compare the performance of the features under session variability. It is sampled at 8 kHz, and we use recordings from 138 speakers. For each speaker, from the first session, the first 20 utterances are used for training and from all the four sessions, the features obtained from the next four utterances are together used for testing.

The feature vectors extracted from the training data are scaled to unit norm. Using these features, we model each speaker by 32-component Gaussian mixture with diagonal covariance matrix, denoted by $S_\theta$. Here, $\theta$ denotes the model parameters (mean $\mu$, covariance $\Sigma$ and weights $w$ of mixture components). The same GMM configuration is used for all the databases, and for MFCC features also. The test data is classified as belonging to the speaker having the maximum per-sample average log-likelihood, obtained as,

$$L(\theta|x) = \frac{1}{n}\sum_{i=1}^{n}\log(S_\theta(x_i)) \quad (1)$$

where, $S_\theta(x_i) = \sum_{j=1}^{k} w_j f(x_i|\mu_j,\Sigma_j)$; $k$ (=32) is the number of mixture components and $n$ is the number of feature vectors $[x = (x_1, x_2, x_3, ...., x_n)]$ available in the test data. The modelling and likelihood estimation are performed using scikit-learn (Pedregosa et al., 2011).

The performance of PS-DCT features is compared with those of the existing glottal based features and 13-dimensional MFCCs. MFCCs are computed only from the voiced segments of the speech signal for all the databases, so that the comparison is fair. A frame length of 30 ms with Hanning window and a frame shift of 10 ms are used for computing the MFCC features. The GMM described above is used for the MFCC features also. The other features compared are: (i) the deterministic plus stochastic model (DSM) of the residual signal, proposed by Drugman and Dutoit (Drugman and Dutoit, 2012), which they used for SID. (ii) DCT of the integrated linear prediction residual (ILPR), proposed in (Abhiram et al., 2015), where ILPR is used as a voice source estimate. The results reported for these two features are taken from the literature.

## 3.2 Speaker Verification (SV) Studies

An i-vector based speaker verification system has been implemented using Microsoft Identity Toolbox (Sadjadi and Omid, 2013)for this work. Entire TIMIT database and a subset of Mandarin corpus (7000 utterances from 700 speakers) are used as development data for obtaining universal background model (UBM) of 256 mixtures and total variability subspace (T-matrix of 400 columns). Totally, we are using approximately 12 hours of data from 542 female and 788 male speakers. From the YOHO database, we have used recordings from four sessions of each of 138 speakers for training their speaker models and evaluating the performance. Verification trials consist of all possible model-test combinations, resulting in a total of 19,044 ($138 \times 138$) trials (138 target versus 18,906 impostor trials). From the first session, we have used the first 20 utterances from each speaker as enrollment data for training speaker models and from all the four sessions, the data obtained from the next four utterances are used for testing. From Mandarin corpus, 24,025 trials (155 target versus 23,870 impostor trials) from 155 speakers are used for testing. 10 utterances from each speaker are used for enrollment and data from four utterances are used for testing.

In i-vector based SV systems, both the training and test segments are represented by i-vectors. The dimensionality of the i-vectors is reduced using 200-dimensional linear discriminant analysis (LDA) to remove channel directions in order to increase the discrimination between speaker subspaces. The Baum-Welch statistics (Dehak et al., 2011) are computed from the training and test feature vectors. Using this statistics along with T-matrix, we compute the train and test i-vectors. After mean and length normalization (Garcia-Romero and Espy-Wilson, 2011), the i-vectors are modelled via a generative factor analysis approach called the probabilistic LDA (PLDA). After that, a whitening transformation is applied, which is learned from the i-vectors of the development set (Sadjadi and Omid, 2013). Finally, a linear strategy is used for scoring the verification trials (Sadjadi and Omid, 2013), which computes the log-likelihood ratio of same to different speakers' hypotheses.

Separate SV experiments are performed using MFCCs, PS-DCT and their score level combination. For more details about the implemented SV system, one can refer to Microsoft Identity Toolbox (Sadjadi and Omid, 2013). We have extracted 13 MFCCs with their delta and delta-delta coefficients to form 39-dimensional feature vectors. Cepstral mean and variance normalization (CMVN) is used for further processing. Since the dimension of PS-DCT depends on the sampling rate, we have extracted PS-DCT features after re-sampling all the utterances to 8 kHz. 56-dimensional PS-DCT features are obtained as mentioned in Sec. 2. Since PS-DCT is not in the cepstral domain, we have not applied CMVN. Instead, we have scaled the feature vectors individually to unit norm (feature vector length) for post-processing.

Table 2: Identification accuracies (in %) on 168 speakers from the TIMIT database, 855 from the Chinese Mandarin corpus and 138 from the YOHO database for four different recording sessions.

| Datasets | DCTILPR | DSM | PS-DCT | MFCC |
|---|---|---|---|---|
| **TIMIT** | 94.6 | 98.0 | 99.4 | 99.4 |
| **Mandarin** | NA | NA | 96.1 | 98.1 |
| **YOHO (Same session)** | 100 | 99.7 | 100 | 100 |
| **YOHO (1 session later)** | 80.4 | 69.3 | 86.9 | 96.3 |
| **YOHO (2 session later)** | 73.9 | 64.3 | 92.0 | 97.8 |
| **YOHO (3 session later)** | 72.5 | 58.7 | 80.4 | 91.3 |

# 4 RESULTS AND DISCUSSION

## 4.1 Results of Speaker Identification Experiments

Table 2 shows that the PS-DCT, MFCC and DSM features perform well on TIMIT dataset. To confirm our hypothesis that the pitch-synchronous analysis offers an advantage, we also extracted DCT features using the traditional fixed length windows. With a frame length of 20 ms and a frame shift of 10 ms, we obtained a maximum accuracy of 75% using 30 coefficients on TIMIT dataset. However, the best accuracy of 88.6% is obtained with a frame length of 10 ms, shift of 5 ms and 50 coefficients. Thus, the performance of PS-DCT at 99.4% is far superior, and hence, for the rest of the databases, we experimented only with PS-DCT. On YOHO dataset, PS-DCT achieves better results than those of DCTILPR and DSM, though it also suffers from session variability. On Mandarin corpus, PS-DCT performs well (96.1%), but not better than MFCCs. By observing the performance of PS-DCT on various datasets, we can see that it has potential to capture speaker specific, discriminative information.

## 4.2 Results of Speaker Verification Experiments

In real world deployment, the SV systems need a minimum duration of test data for robust performance. In some applications, this may not be possible, which leads to poor system performance. Hence, it is of interest to build a SV system that can perform better even when the data available for testing is limited. Later, we see that PS-DCT-based SV system can boost the performance of MFCC-based SV system significantly under limited test data condition.

Table 3 shows the results obtained on YOHO database and Mandarin corpus for 19,044 and 24,055 trials, respectively, using both MFCC and PS-DCT features. Table 3 shows that when the duration of the test data is low ($<$ 4 sec), the EER is quite higher than the case where the duration of the test data is higher ($>$ 4 sec) for both the features. Here, PS-DCT alone does not perform as good as MFCC. In an earlier study (Das et al., 2014), it has been shown that fusing MFCCs and other feature based classifiers using a convex combination improves the performance of the speaker verification system significantly. So, here also we have applied a convex combination at the score level to get the score for the combined system as follows:

$$S_{\text{COMB}} = \lambda S_{\text{PS-DCT}} + (1-\lambda) S_{\text{MFCC}} \qquad (2)$$

where, $S_{\text{COMB}}$, $S_{\text{PS-DCT}}$ and $S_{\text{MFCC}}$ are the scores obtained from the combined, PS-DCT and MFCC based SV systems, respectively. $\lambda$ is a scalar between 0 and 1.

Using the combined score $S_{\text{COMB}}$ improves the performance of the SV system significantly for both the databases, specifically when the duration of the test data is limited ($<$ 3 sec). The best value of $\lambda_{opt}$ is experimentally determined to be 0.4 for both the databases. Table 4 shows the performance of the combined system for both YOHO and Mandarin databases along with the absolute improvement in EER (in %). For YOHO database, the performance of the combined system is higher for all the sessions for all the durations of the test data. The performance improvement is significantly high (up to 5.8% in absolute terms) for very limited duration of the test data (1.5 sec). The same is true for Mandarin corpus also, where the improvement is up to 3.2% for test data of 2 sec. In both these cases, there is sufficient training data. Table 4 shows that the absolute improvement in EER over the MFCC system increases as the dura-

Table 3: Speaker verification performance of MFCC-based (PS-DCT-based) i-vector system on YOHO database and Mandarin corpus, for different durations of the test data. Number of trials for YOHO: 19,044; Mandarin: 24,025.

| **YOHO database (equal error rate (EER) in %)** | | | | | **Mandarin corpus** | |
|---|---|---|---|---|---|---|
| **Test data (approx.)** | **Same session** | **1 session later** | **2 sessions later** | **3 sessions later** | **Test data (approx.)** | **EER (%)** |
| 1.5 sec | 11.5 (12.2) | 23.1 (28.2) | 18.8 (22.8) | 19.5 (25.3) | 2 sec | 12.2 (16.0) |
| 3 sec | 5.2 (7.2) | 14.0 (20.2) | 9.4 (18.0) | 8.7 (18.0) | 4 sec | 5.1 (9.6) |
| 6 sec | 1.4 (4.3) | 7.2 (11.0) | 7.1 (12.3) | 6.5 (16.6) | 8 sec | 2.6 (5.5) |
| 9 sec | 1.0 (2.8) | 4.3 (9.1) | 6.5 (10.8) | 5.7 (14.4) | 12 sec | 2.4 (4.0) |

Table 4: Performance of the combined i-vector speaker verification system (EER in %) on YOHO and Mandarin databases for different durations of the test data (with $\lambda_{opt} = 0.4$). The values within brackets give the absolute improvement in EER over the MFCC-based system.

| **YOHO database (equal error rate (EER) in %)** | | | | | **Mandarin corpus** | |
|---|---|---|---|---|---|---|
| **Test data (approx.)** | **Same session** | **1 session later** | **2 sessions later** | **3 sessions later** | **Test data (approx.)** | **EER (%)** |
| 1.5 sec | 6.4 (5.1) | 17.3 **(5.8)** | 14.1 (4.7) | 16.4 (3.1) | 2 sec | 9.0 **(3.2)** |
| 3 sec | 2.1 (3.1) | 11.4 (2.6) | 6.0 (3.4 | 7.9 (0.8) | 4 sec | 3.2 (1.9) |
| 6 sec | 0.1 (1.3) | 5.0 (2.2) | 4.1 (3) | 5.7 (0.8) | 8 sec | 1.2 (1.4) |
| 9 sec | 0.0 (1.1) | 3.6 (0.7) | 4.1 (2.4) | 5.0 (0.7) | 12 sec | 1.2 (1.2) |

tion of the test data reduces. Thus, PS-DCT does capture speaker-specific information that is supplementary to the information captured by MFCCs and the performance gain of the combined system is higher for limited durations of the test data ($< 3$ sec). Since PS-DCT is directly computed on the speech signal, it captures the voice source information in addition to the vocal tract information, and hence is able to supplement the performance of MFCCs in the speaker verification task.

# 5 CONCLUSION

Based on experiments performed on different datasets, we show that the PS-DCT, derived from the voiced part of the speech signal, is an effective feature for SID/SV tasks (see Tables 2, 3, and 4). The proposed PS-DCT can capture the fine level differences in the 'time-domain waveform shape' of a particular voiced phone from different speakers (see Fig. 3).

# REFERENCES


Abhiram, B., Ramakrishnan, A. G., and Prasanna, S. M. (2015). Voice source characterization using pitch synchronous discrete cosine transform for speaker identification. *J. Acoust. Soc. Am.*

Campbell, J. and Higgins, A. (1994). YOHO speaker verification.

Das, R. K., Abhiram, B., Prasanna, S., and Ramakrishnan, A. (2014). Combining source and system information for limited data speaker verification. *Proc. 15th Annual Conf. of the International Speech Communication Association.*

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Process.*, 28(4):357–366.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 19(4):788–798.

Drugman, T. and Dutoit, T. (2012). The deterministic plus stochastic model of the residual signal and its applications. *IEEE Trans. Audio, Speech, Lang. Process.*, pages 968–981.

Garcia-Romero, D. and Espy-Wilson, C. Y. (2011). Analysis of i-vector length normalization in speaker recognition systems. *Twelfth Annual Conference of the International Speech Communication Association*.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Technical Report*, 93.

Gudnason, J. and Brookes, M. (2008). Voice source cepstrum coefficients for speaker identification. *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. ICASSP* ., pages 4821–4824.

Mathews, M. V., Miller, J. E., and Jr., E. E. D. (1961). Pitch synchronous analysis of voiced sounds. *J. Acoust. Soc. Am.*, 33.

Nakagawa, S., Wang, L., and Ohtsuka, S. (2012). Speaker identification and verification by combining MFCC and phase information. *IEEE Trans. Audio, Speech, Lang. Process.*, 20(4):1085–1095.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Prasanna, S. M., Gupta, C. S., and Yegnanarayana, B. (2006). Extraction of speaker-specific excitation information from linear prediction residual of speech. *Speech Communication*, 48(10):1243–1261.

Prathosh, A. P., Ananthapadmanabha, T. V., and Ramakrishnan, A. G. (2013). Epoch extraction based on integrated linear prediction residual using plosion index. *IEEE Trans. Audio, Speech, Lang. Process.*, 21(12):2471 – 2480.

Sadjadi and Omid, S. (2013). *MSR identity toolbox v1.0: A MATLAB toolbox for speaker recognition research*. Microsoft research technical report.