Voice source characterization using pitch synchronous discrete cosine transform for speaker identification

A.G. Ramakrishnan and B. Abhiram
Department of Electrical Engineering, Indian Institute of Science, Bangalore, India 560012
ramkiag@ee.iisc.ernet.in, abhiram1989@gmail.com

S.R. Mahadeva Prasanna
Department of Electrical and Electronics Engineering, I.I.T.-Guwahati, India 781039
prasanna@iitg.ernet.in

Abstract: A characterization of the voice source (VS) signal by the pitch synchronous (PS) discrete cosine transform (DCT) is proposed. With the integrated linear prediction residual (ILPR) as the VS estimate, the PS DCT of the ILPR is evaluated as a feature vector for speaker identification (SID). On TIMIT and YOHO databases, using a Gaussian mixture model (GMM)-based classifier, it performs on par with existing VS-based features. On the NIST 2003 database, fusion with a GMM-based classifier using MFCC features improves the identification accuracy by 12% in absolute terms, proving that the proposed characterization has good promise as a feature for SID studies.

PACS numbers: 43.72.Ar, 43.72.Pf, 43.70.Bk

1. Introduction

In the production of voiced speech, the derivative of the glottal flow is called the voice source (VS) signal. In the source-filter model, speech is modeled as the output of the vocal tract filter, excited by the VS. The VS pulse synthesized by models[1] and the VS estimated by inverse filtering the speech signal have been used as the source signal for speech synthesis[2]. The VS estimate is also used for analyzing pathological voices[3] and features extracted from its shape, for speaker identification (SID)[4;5;6]. Further, studies like[7] show that the VS pulse shape influences the perceived voice quality. Time-domain models have been proposed to characterize the VS pulse[8]. The spectrum of the VS pulse has also been parameterized[9]. In other works, the samples of the VS estimate have been directly used[10] or its frequency or cepstral-domain representation[4;5].

The objective of this study is to propose an alternate way of characterizing the VS, and to evaluate it as a feature for SID. Thus, the focus is not on the speaker modeling and classification. The discrete cosine transform (DCT) is a reversible transformation, with an excellent energy compaction property, and hence has the ability to capture the time-domain pulse shape of the VS within its first few coefficients. Since the pulse shape of the VS has been successfully exploited for SID[4;5], its DCT is explored as an alternate characterization of the VS for SID.

2. Discrete cosine transform of the integrated linear prediction residual

In earlier studies like [4;5], the closed-phase covariance technique of linear prediction (LP) analysis[11] was used to obtain the VS estimate. For proper estimation, this technique requires the glottal cycle to have a sufficiently long closed-phase, which is not the case in breathy phonation, where the vocal folds do not close fully[12]. To avoid such dependence on the speech signal, we use the integrated linear prediction residual (ILPR)[7] as the VS estimate, since it only involves estimating the LP coefficients from the pre-emphasized speech signal and using them to inverse filter the non-pre-emphasized speech signal, without the need to estimate the closed-phase prior to inverse filtering.

The vowel /a/ shown in Fig. 1(a) is synthesized from the VS pulse (dotted line in Fig. 1(b)) simulated using the model in [13]. The ILPR (solid line), also shown in Fig. 1(b), bears a close resemblance to the original VS pulse, except for small ripples in the closed phase of the glottal cycle, resulting from improper formant cancellation,

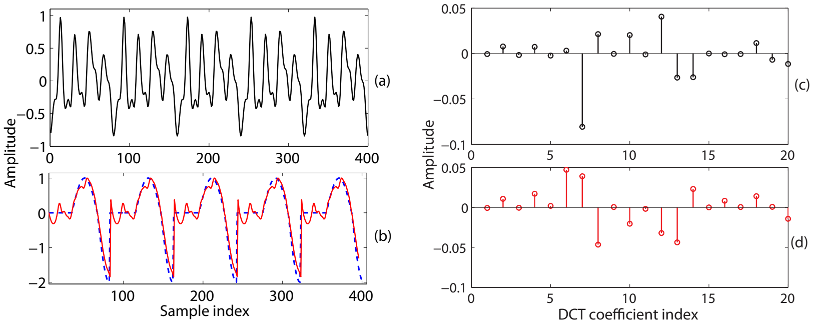

Fig. 1. Shift variance of the DCT: (a) A segment of synthetic vowel /a/; (b) The ILPR (solid line) closely resembles the VS pulse train (dotted line) used for synthesis; (c) The first 20 coefficients of the pitch synchronous (PS) DCT of the ILPR in Fig. 1(b); (d) The first 20 coefficients of the PS DCT of the circularly shifted version of the ILPR in Fig. 1(b) are different from those in (c).

2.1. Pitch synchronous discrete cosine transform and the number of coefficients

It is desirable that the representation of the VS is independent of shift and scale changes. However, the DCT is shift-variant and Figs. 1(c) and (d) show that the DCT coefficients of a segment of the ILPR shown in Fig. 1(b) and its circularly shifted version are quite different. This problem is avoided if the DCT is obtained pitch synchronously, using pitch marks. Pitch synchronous DCT has been used for pitch modification[14]. The DCT has also been demonstrated to be a good feature extractor for a few recognition problems[15].

Since the VS is by nature low-pass[13], it is not necessary to retain the DCT coefficients corresponding to high frequencies. Consider the DCT of an ILPR sequence of length N equal to one pitch period (typically 50-160 samples for $f_s = 16\ kHz$). Let the first M DCT coefficients (excluding the 0^{th}) be retained. To choose M, a segment of the ILPR from a TIMIT utterance is reconstructed using the N-point inverse DCT by varying M (forcing the remaining $N - M$ coefficients to zero). Fig. 2 shows the reconstructed ILPR pulses (for $M = 12,\ 24$ and 50) overlaid on the original. As M increases, the re-

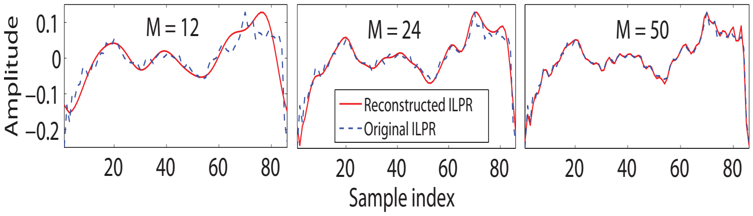

Fig. 2. A single pitch period of the ILPR (dotted line) of a segment of a TIMIT utterance and its reconstruction (solid line) from DCT truncated to 12, 24 and 50 coefficients. The gross shape of the ILPR is captured by only 12 of the more than 80 DCT coefficients.

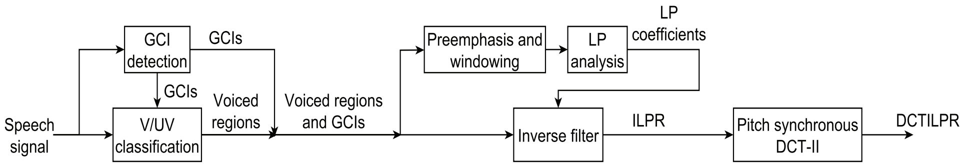

Fig. 3. Block diagram of the method to obtain DCT of ILPR.

construction gets closer to the original. This behaviour is similar to that of the principal components of the VS proposed in[16]. The gross shape of the ILPR is captured by only 12 coefficients. However, this is just an example, and we require statistics from an SID experiment to determine the optimal value of M.

The block diagram to obtain the DCT-based characterization (henceforth referred as the DCTILPR) is shown in Fig. 3. An algorithm for glottal closure instant (GCI) detection[17] is used for pitch marking. Since only the voiced regions are of interest, a voiced/unvoiced (V/UV) classification scheme based on maximum normalized cross correlation is used to retain only the voiced regions, as in[18]. The ILPR is obtained by inverse filtering three successive pitch periods, retaining only the middle period of the output, and repeating the process by shifting the analysis frame by one pitch period till the entire voiced speech segment is traversed. A first order difference filter is used for pre-emphasis and a Hanning window, for LP analysis. Each ILPR cycle is normalized by its positive peak amplitude, before applying the DCT-II pitch synchronously.

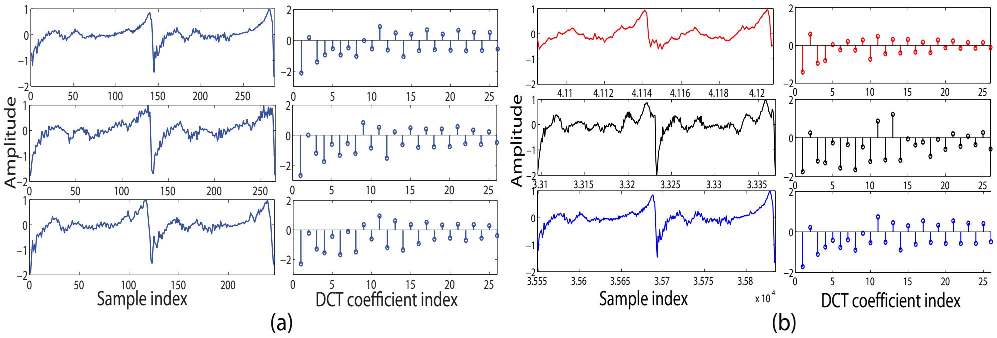

Fig. 4. (a) Two periods of ILPR and the PS DCT (of one period) from different phonetic contexts in voiced segments from three different utterances of a single TIMIT speaker. The shapes of the ILPRs are similar and so are their DCTs. (b) ILPR and PS DCT from vowel /a/ in the word ‘wash’ for three TIMIT speakers. The ILPR shape, and hence the DCT vectors vary across speakers.

2.2. Speaker variability: modeling using Gaussian mixture models

Fig. 4(a) illustrates that the ILPRs from three different utterances of the same speaker are similar, and the corresponding DCT coefficients show a similar trend, though not identical. On the other hand, Fig. 4(b) shows that different speakers have distinct ILPR waveform shapes and DCT vectors. Thus, the DCTILPR captures the ILPR shape and has less intra-speaker variability and more inter-speaker variability, which is a characteristic of a good speaker-specific feature.

Even though the ILPR shape is similar in most regions for a single speaker, it is different in some regions. This is due to improper formant cancellation during inverse filtering, leading to changes in ILPR shape for different phones, and also due to the different phonation types[19]. We use Gaussian mixture models (GMMs)[20] as speaker models to capture the changes in the feature distribution from speaker to speaker.

3. Evaluation of the proposed feature for speaker identification

Three standard databases are used to evaluate the efficacy of the DCTILPR as a feature vector for SID: (1) The TIMIT test set[21] with 168-speaker data is used for comparison with existing VS-based features for SID[22;5;4]. Utterances 3 to 10 are used as train data and 1 and 2 as test data for each speaker. (2) The YOHO database[23] with data from 4 different recording sessions for each of 138 speakers is used to study the session variability of the DCTILPR. Utterances 1 to 20 are used for training and 21 to 24 for testing, for each session of each speaker. (3) The NIST 2003 database[24] with 356-speaker, multiple cellular phone speech (110 s of train data and 60 s of test data per speaker) is used to study the handset variability of the DCTILPR. A Gaussian mixture with 16 components and diagonal covariance matrices is used to model each speaker and the decision is based on maximum likelihood.

The performance of DCTILPR is compared with those of the following studies on VS-based features for SID: (1) Plumpe et. al.[4], with a time-domain parameterization (TDVS) of the VS estimate as the feature, and a GMM-based classifier; (2) Gudnason and Brookes[5], with the voice source cepstral coefficients (VSCC) as the features, and a GMM-based classifier; and (3) Drugman and Dutoit[22], with the speaker-dependent waveforms of the deterministic plus stochastic model (DSM) of the LP residual as the features, and a distance metric for classification.

	M	4	8	12	16	20	24	28	32	36
SID accuracy (%)	TIMIT	75.6	86.8	92.8	93.5	93.5	**94.6**	92.9	92.3	94.0
	YOHO	52.9	74.6	78.8	79.0	79.0	**80.4**	79.0	76.8	75.4

Table 1. SID accuracy versus M on TIMIT and YOHO databases

3.1. Results and discussion

Table 1 shows the performance variation with M on TIMIT and YOHO (training data – session 1 and test data – session 2). It shows that the first 12 DCT coefficients are the most effective for SID, which implies that the gross shape of the VS plays a significant role in distinguishing between speakers. Also, the performance drop for $M > 24$ means that the very fine structure of the VS does not capture the features distinguishing the speakers. Since $M = 24$ gives the maximum accuracy, the first 24 coefficients are taken as the feature vector for our experiments.

Table 2(a) shows that the performance of the DCTILPR on TIMIT is comparable to that of the VSCC, but the DSM outperforms both with only 3 misidentifications. However, Table 2(b) shows that, on the YOHO database, the DSM, which is the best performing feature in[22], is outperformed by the DCTILPR in all the 4 sessions. In Table 2(b), I position means correct classification, and II and III positions mean that the true speakers correspond to the speaker models having the second and third highest likelihoods for the given test data, respectively. With both the features, the performance decreases from session to session, and more speakers are pushed to II and III positions. Thus there is session variability in both the features, which needs to be alleviated to be able to use them in a practical setting. The difference between the performances of the DCTILPR on the TIMIT and YOHO databases may be due to the availability of less training (24s) and test data (6s) in TIMIT.

On the TIMIT database, the LF model parameters are obtained as described in[4]. From both the LF model parameters and the DCTILPR coefficients, the ILPR is reconstructed and the ratio of the energy of the reconstruction error to the energy of

Feature	# misidentifications	SID accuracy
TDVS	48	71.4%
VSCC	9	94.6%
DCTILPR	9	94.6%
DSM	**3**	**98.0%**

(a)

Session details	In I position	In II position	In III position
Same session	**100.0%** (99.7%)	0% (0.3%)	0% (0%)
1 session later	**80.4%** (69.3%)	3.6% (7.9%)	2.9% (5.2%)
2 sessions later	**73.9%** (64.3%)	2.9% (8.8%)	5.1% (4.6%)
3 sessions later	**72.5%** (58.7%)	5.1% (11.8%)	3.7% (4.4%)

(b)

Table 2. (a) Performance comparison on TIMIT database; (b) Percentage of speakers classified using DCTILPR (using DSM) with test data from different recording sessions in YOHO database

Condition	DCTILPR	MFCC	Classifier fusion	
			DCTILPR+MFCC [$\alpha.L_C + (1-\alpha).L_M$]	LF+MFCC [$\alpha.L_F + (1-\alpha).L_M$]
Same handset	71.8%	72.7%	Max. accuracy = **84.5%**	Max. accuracy = 75.45%
Different handset	18.2%	40.0%	Max. accuracy = 40.9%	Max. accuracy = 40.0%

Table 3. SID accuracy on the NIST 2003 database under the same and different handset conditions for DCTILPR, MFCC and classifier fusion.

the original ILPR is noted. The mean (over multiple pitch periods) reconstruction error energy with the LF model is 0.42, while that with the DCTILPR (24 coefficients) is 0.23. This shows that the DCTILPR captures the finer details of the VS pulse, while the LF model does not. This may be the reason for the better SID performance of the DCTILPR.

On the entire NIST 2003 database, the DCTILPR gives a low SID accuracy of 16.5%, probably due to the handset variability. To test this, a 110-speaker subset of the database is considered having training and test data from the same and also different handsets. The performance of the DCTILPR is compared with that of the MFCC, computed only from the voiced segments, to be consistent with the DCTILPR. Table 3 shows that the SID accuracy for the same handset is around 72% for both the features. However, when the handset is different, it drops drastically to 18% with the DCTILPR and to 40% with the MFCC. Thus there is handset variability in both the features, but the DCTILPR suffers more than the MFCC. This is mostly because the MFCC captures only the magnitude while the DCTILPR also captures the phase, causing it to suffer more from phase response variations between different microphones and channels.

However, fusing the classifiers by combining their likelihoods L_C (DCTILPR) and L_M (MFCC) linearly as $\alpha.L_C + (1-\alpha).L_M$, where α is a scalar varying from 0 to 1, improves the absolute SID accuracy by 12% over the one using only the MFCC, showing that the DCTILPR can significantly supplement the MFCC in SID systems, in the same handset case. This is because the MFCCs mainly capture the vocal tract characteristics, whereas the DCTILPR captures the VS characteristics. This result is similar to that in[6] which reports a 7% improvement in SID accuracy after combining VS-based features with MFCCs. In the different handset case, there is no improvement with classifier fusion, probably due to the large handset variability suffered by both the features.

Combining the likelihoods L_F and L_M of the classifiers trained with the LF model parameters and the MFCCs shows only a 2.73% improvement in the same handset case. This may be because, in most cases in the NIST 2003 database, the ILPR shape is different from the LF model shape, which might be due to channel noise and filtering effects during cellular communication. When the ILPR is reconstructed, the mean of the normalized reconstruction error energy using DCTILPR is 0.28, and 0.63 using the LF model. Thus the rigid time-domain LF model is not able to adapt to changes in pulse shape, and hence characterizing the VS pulse using the reversible DCT is a better option.

4. Conclusion

The results show that the DCTILPR has a good promise as a feature for SID studies. It is inferred that the gross shape of the VS (captured by the 12 DCT coefficients) is crucial in distinguishing speakers, and the fine structure (captured by the coeffficients higher than the 24th) may increase confusion among speakers. The DCTILPR suffers from session and handset variabilities, which may be compensated by techniques like within-class covariance normalization[25] to deploy it in a practical scenario. A fusion of classifiers trained using the DCTILPR and the MFCC improves the performance, showing that the DCTILPR can supplement the MFCC in SID systems. A comparison of reconstruction error energy shows that characterizing the VS by the reversible DCT is better than fitting a rigid time-domain model to it.

Acknowledgments

The authors thank Dr. T.V. Ananthapadmanabha for sharing his wealth of knowledge on the voice source and the ILPR, and Prathosh for the stimulating discussions.

References and links

[1] D.H. Klatt. Review of text–to–speech conversion for English. *J. Acoust. Soc. Am.*, 82(3):737–793, 1987.

[2] P. Alku, H. Tiitinen, and R. Naatanen. A method for generating natural–sounding speech stimuli for cognitive brain research. *Clinical Neurophysiology*, 110:1329–1333, 1999.

[3] Y. Koike and J. Markel. Application of inverse filtering for detecting laryngeal pathology. *Annals of otology, rhinology and laryngology*, 84:117–124, 1975.

[4] M.D. Plumpe, T.F. Quatieri, and D.A. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. Speech Audio Process.*, 7:569–586, 1999.

[5] J. Gudnason and M. Brookes. Voice source cepstrum coefficients for speaker identification. In *Proc. ICASSP*, pages 4821–4824, 2008.

[6] J. Wang and M. T. Johnson. Physiologically-motivated feature extraction for speaker identification. In *Proc. ICASSP*, pages 1690–1694, 2014.

[7] T.V. Ananthapadmanabha. Acoustic factors determining perceived voice quality. In *Vocal fold physiology: voice quality control*, pages 113–126. O.Fujimura and M. Hirano, Eds., San Diego, Cal.: Singular publishing group, ch. 7, 1995.

[8] G. Fant, J. Liljencrants, and Q. Lin. A four–parameter model of glottal flow. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 26:1–13, 1985.

[9] P. Alku, H. Strik, and E. Vilkman. Parabolic spectral parameter – a new method for quantification of the glottal flow. *Speech Commun.*, 22:67–79, 1997.

[10] S.R.M. Prasanna, C.S. Gupta, and B. Yegnanarayana. Extraction of speaker–specific excitation information from linear prediction residual of speech. *Speech Commun.*, 48:1243–1261, 2006.

[11] D.Y. Wong, J.D. Markel, and A.H. Gray Jr. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Trans. Acoust., Speech. Signal Process.*, 27:350–355, 1979.

[12] P. Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Commun.*, 11(2-3):109–118, 1992.

[13] T.V. Ananthapadmanabha. Acoustic analysis of voice source dynamics. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 25(2–3):1–24, 1984.

[14] R. Muralishankar, A.G. Ramakrishnan, and P. Prathibha. Modification of pitch using DCT in the source domain. *Speech Commun.*, 42(2):143–154, 2004.

[15] P.B. Pati and A.G. Ramakrishnan. Word level multi–script identification. *Pattern Recognition Letters*, 29:1218–1229, 2008.

[16] J. Gudnason, M.R.P. Thomas, D.P.W. Ellis, and P.A. Naylor. Data-driven voice source waveform analysis and synthesis. *Speech Commun.*, 54(2):199–211, 2012.

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

[18] T.V. Ananthapadmanabha, A.P. Prathosh, and A.G. Ramakrishnan. Detection of closure-burst transitions of stops and affricates in continuous speech using plosion index. *J. Acoust. Soc. Am.*, 135(1): 460–471, 2014.

[19] D.G. Childers and C. Ahn. Modeling the glottal volume velocity waveform for three voice types. *J. Acoust. Soc. Am.*, 97(1):505–519, 1995.

[20] D.A. Reynolds and R.C. Rose. Robust text–independent speaker identification using gaussian mixture speaker models. *IEEE Trans. Speech Audio Process.*, 3(1):72–83, 1995.

[21] W. Fisher, G. Doddington, and K. Goudie–Marshall. The DARPA speech recognition research database: Specifications and status. In *Proc. DARPA Workshop on Speech Recognition*, pages 93–99, 1986.

[22] T. Drugman and T. Dutoit. The deterministic plus stochastic model of the residual signal and its applications. *IEEE Trans. Audio, Speech, Lang. Process.*, 20:968–981, 2012.

[23] J. Campbell. Testing with the YOHO CD–ROM voice verification corpus. In *Proc. ICASSP*, pages 341–344, 1995.

[24] NIST Multimodal Information Group. 2003 NIST Speaker Recognition Evaluation, Linguistic Data Consortium, Philadelphia.

[25] A. Hatch, S. Kajarekar, and A. Stolcke. Within-class covariance normalization for SVM-based speaker recognition. In *Proc. Intl. Conf. Spoken Lang. Process.*, 2006.