Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction *

Ville Vestman[1], Dhananjaya Gowda[2], Md Sahidullah[1],
Paavo Alku[3], Tomi Kinnunen[1]

[1] *School of Computing, University of Eastern Finland, Finland*
[2] *DMC R&D Center, Samsung Electronics, Seoul, Korea*
[3] *Department of Signal Processing and Acoustics, Aalto University, Finland*

Abstract

From the available biometric technologies, automatic speaker recognition is one of the most convenient and accessible ones due to abundance of mobile devices equipped with a microphone, allowing users to be authenticated across multiple environments and devices. Speaker recognition also finds use in forensics and surveillance. Due to the acoustic mismatch induced by varied environments and devices of the same speaker, leading to increased number of identification errors, much of the research focuses on compensating for such technology-induced variations, especially using machine learning at the statistical back-end. Another much less studied but at least as detrimental source of acoustic variation, however, arises from mismatched speaking styles induced by the speaker, leading to a substantial performance drop in recognition accuracy. This is a major problem especially in forensics where perpetrators may purposefully disguise their identity by varying their speaking style. We focus on one of the most commonly used ways of disguising one's speaker identity, namely, whispering. We approach the problem of normal-whisper acoustic mismatch compensation from the viewpoint of robust feature extraction. Since whispered speech is intelligible, yet a

*This work contains limited portions of [1].

This is the accepted manuscript of an article published in Speech Communication.
Link to the formal publication: https://doi.org/10.1016/j.specom.2018.02.009

low-intensity signal and therefore prone to extrinsic distortions, we take advantage of robust, long-term speech analysis methods that utilize slow articulatory movements in speech production. In specific, we address the problem using a novel method, *frequency-domain linear prediction with time-varying linear prediction* (FDLP-TVLP), which is an extension of the *2-dimensional autoregressive* (2DAR) model that allows vocal tract filter parameters to be time-varying, rather than piecewise constant as in classic short-term speech analysis. Our speaker recognition experiments on the whisper subset of the CHAINS corpus indicate that when tested in normal-whisper mismatched conditions, the proposed FDLP-TVLP features improve speaker recognition performance by 7–10% over standard MFCC features in relative terms. We further observe that the proposed FDLP-TVLP features perform better than the FDLP and 2DAR methods for whispered speech.

Keywords: Speaker recognition, speaking style mismatch, disguise, whisper, 2-dimensional autoregression (2D-AR), time-varying linear prediction (TVLP)

1. Introduction

Research in *automatic speaker recognition* [2] has focused increasingly on enhancing robustness in adverse conditions induced by background noise, reverberation and low-quality recordings. Many approaches have been studied to tackle these challenges, one of the most successful being the *i-vector* technology [3] used jointly with the *probabilistic discriminant analysis* (PLDA) back-end [4, 5]. In addition to the new utterance level features (i-vectors) and back-ends, improvements have been achieved in the first part of the speech processing chain by developing *robust acoustic features* [6, 7, 8]. Recent advances in both topics have brought the performance of speaker recognition systems closer to the level expected in applications such as forensics, surveillance, and authentication.

In addition to the environment-related and technology-related acoustic variations, another major, yet much less studied problem arises from *within-speaker* variations caused by differences in speaking styles. Changes in the speaking style occur, for instance, when the speaker shouts [8] or whispers [9]. Current recognition systems that are typically trained with speech of normal speaking style can tolerate only small changes in the speaking style, and reliable speaker recognition has turned out to be very challenging if the mode of speaking changes considerably from normal [9, 10, 8]. Speaking style mismatched speaker recognition has applications for example in forensics, where recorded audio excerpts can be used as evidence. For instance, a crime might be committed in an agitated state of mind, leading to shouting or screaming. Similarly, a perpetrator might deliberately disguise his or her identity in order to avoid being identified [11]. In [12], whispering was found to be the most common way of changing the mode of speech production to disguise the speaker identity. Furthermore, [12] and [13] report that disguise is commonly found in criminal action, especially in blackmailing cases. Whispering can also be used in public places to prevent others from hearing private information or to avoid disturbing others in places where silent behavior is expected. Conversely, people tend to use loud, high-effort voice in noisy environments in order to make their speech more intelligible in background noise. The tendency of the speaker to change his or her speaking style in noisy environments is known as the *Lombard effect* [14].

As described above, various speaking styles are expected to be encountered in real-world speaker recognition applications. This imposes a considerable challenge to the existing systems whose performance has been shown to drastically decline due to changes in the speaking style [9, 10, 8]. In order to improve the performance of speaker recognition in real-world scenarios including various speaking styles, the current study focuses on a specific style of speaking, whispering, which differs vastly from normal speech in its acoustic properties. In addition to being lower in intensity, whispered speech lacks the vibration of the vocal folds (even in case of voiced sounds such as vowels) when the sound excitation is generated in the larynx [15]. In addition to the absence of the vocal fold vibration, it has been observed that whispered vowels show an upward shift in formant frequencies when compared to vowels of the normal speaking style and that whispered consonants show increased spectral flatness [15].

In principle, suppressing the unwanted within-speaker variations induced by speaking style mismatch could be addressed using statistical back-end methods. In fact, most modern speaker recognition back-ends (as reviewed in [16]) include some kind of a within-speaker variation model, intended to quantify the extent of allowed variation in any pair of utterances of the same speaker, before they are considered more likely to have been spoken by different speakers. Dating back to Kenny's pioneering work on *joint factor analysis* (JFA) [17], which later inspired the i-vector paradigm [3], these techniques are realized as various flavors of subspace models where the between-

and within-speaker subspaces are modeled using separate factor loading matrices. To exemplify, the *simplified* PLDA model [18] assumes a Gaussian within-speaker variation model shared across all the speakers, parameterized as a residual covariance matrix. The hyperparameters of such back-end models are trained off-line using, typically, thousands of utterances from hundreds of *development speakers*. In order to adopt these back-ends for explicit style variation compensation, a corpus is needed that contains, per each development speaker, utterances spoken in various speaking styles. Unfortunately, this kind of speech data is prohibitively expensive and difficult to collect in quantities required by PLDA back-ends. Moreover, to the best of our knowledge, such large corpora are not publicly available at present. For these reasons, and since the present study addresses speaker recognition using short utterances, we adopt instead the classic Gaussian mixture model-universal background model (GMM-UBM) [19] back-end approach which, in fact, produces competitive accuracy — or even surpasses the i-vector based approach [20, 21, 22] — in the duration conditions considered in this study.

Another commonly applied back-end recipe to enhance the speaker recognition accuracy across varied conditions is *multicondition training* [23, 24]. It utilizes data obtained from different conditions to prepare the back-end components to expect different variations of the speech data. Again, however, since data collection for multiple conditions takes lots of resources and usually is not a realistic requirement for speaker enrollment, a common practice is to artificially generate data by, for example, digitally adding noise. In case of variation caused by the speaking style (such as whispering), however, generation of realistic artificial data is not easy. Therefore, the current study focuses on an alternative approach, robust extraction of features, to tackle the deteriorating effect caused by the speaking style variation. Feature extraction, as the first step in the speech processing chain of any speaker recognition system, has a key role as it provides inputs to the back-end that can be based, for instance, on the GMM-UBM [19], i-vectors [3], or *deep neural networks* (DNNs) [25, 26]. Thus, we find it important to develop and study features that show good performance across a wide variety of settings to make speaker recognition systems less dependent on large amounts of training data from different conditions. To this end, we propose using two recent feature extraction methods [7, 1] for whispered speech that have already shown good results in other studies.

Traditionally, most features used in speaker recognition are computed from short-term analysis using frames that span about 25 ms of speech [27].

While this approach is effective in capturing instantaneous acoustical features of the vocal tract, it ignores long-term properties of speech such as prosody. In addition, the traditional short-term analysis is not capable of taking into account articulation variations, and it lacks other possible benefits of longer-term processing including improved robustness against noise and reverberation [7]. These limitations of the traditional short-term analysis are important factors to consider especially when dealing with whispered speech as whispering has lower intensity than normal speech [28], which makes it more prone to extrinsic disturbances, such as additive noise. Whispered speech also tends to show widening of formant bandwidths (see Figure 1), which makes it harder to accurately detect formants. We hypothesize that a better utilization of contextual information observed over long-time frames can be used to improve formant modeling accuracy over standard short-time analysis.

To study feature extraction based on long-term processing, we propose using *2-dimensional autoregressive features* (2DAR) for whispered speech speaker recognition. In the 2DAR scheme, speech is processed in temporal domain before feeding it to the typical short-term feature extraction pipeline. The temporal processing is achieved using *frequency domain linear prediction* (FDLP) [29, 30], a method that produces smoothed, parametric time-domain Hilbert envelopes of the individual frequency subbands. The smoothed, parametric representation of the subband Hilbert envelopes provides robustness against noise and temporal smearing caused by reverberation [7].

As one of our key contributions, we propose a novel modification of the 2DAR processing by replacing conventional *linear prediction* (LP), conducted after FDLP, with *time-varying linear prediction* (TVLP) [1]. In TVLP, the coefficients of the linear predictive filter are not considered to be stationary but they are time-varying (*i.e.* non-stationary) and expressed using basis functions (such as polynomials or trigonometric functions). The type and number of the basis functions can be tuned to control the rate of change of the underlying vocal tract model. As a result of adopting TVLP, linear prediction filter coefficients follow slowly-varying time-continuous contours, modelled by the basis functions. Therefore, the corresponding features are less prone to change abruptly over time, which is a phenomenon that degrades, for example, conventional LP-based features when speech is of low-intensity (as in whispers) or corrupted by noise. To be able to apply TVLP after FDLP, we modify the original TVLP model [31, 32], which assumes raw waveform as an input, to be applicable to spectro-temporal representations produced

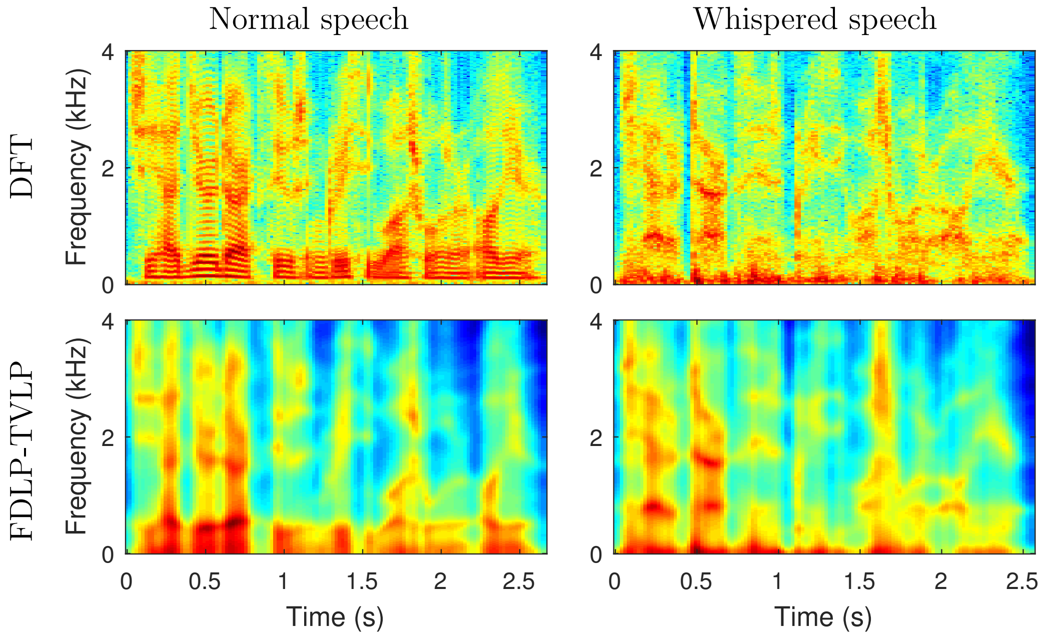


Figure 1: Spectrograms computed using *discrete Fourier transform* (DFT) and the proposed *frequency-domain linear prediction with time-varying linear prediction* (FDLP-TVLP), for the sentence *"She had your dark suit in greasy wash water all year."* uttered by a male speaker in normal and whispered speaking styles. Due to the mismatch in speaking style, the conventional DFT spectrograms between normal and whispered speech differ substantially from each other, while the proposed spectrogram exhibits relatively less variation.

by FDLP.

Next, in Section 2, we present an extensive literature review on automatic speaker recognition from whispered speech and describe available corpora of whispered speech. In Section 3, we discuss the properties of whispered speech and propose a method for automatically comparing formants of normal and whispered speech for a large speech corpus. The proposed method aligns normally spoken sentences with their whispered counterparts by matching the speech content. The method does not require speech transcription with time alignments, so it can be readily applied to any corpus containing parallel data of two speaking styles. In Sections 4, 5, and 6, we extend our preliminary study [1] in many ways both regarding the methodology and experiments. Firstly, we provide a more thorough and self-contained exposition of the used methodologies. Secondly, unlike in [1], we study the choice of the basis functions in TVLP. Thirdly, we verify our early positive findings with a new dataset and a different problem domain ([1] focuses on reverberation robustness rather than speaking style mismatch). Finally, for the comparison, we include a broader set of state-of-the-art reference features including *power-law adjusted linear prediction* (LP-α) [8] and *minimum variance distortionless response* (MVDR) [33] features.

2. Speaker recognition from whispered speech

In Table 1, we have summarized the main characteristics of the speaker recognition studies on whispered speech that we found. We hope that this table gives the interested reader a better understanding of the studies on the topic. For those conducting their own research on whispered speech, we also recommend [34] to get more insight of available speech corpora containing whispered speech.

The performance numbers in the table indicate that the task at hand is a rather difficult one. In the identification task, the accuracy can be as low as 50 per cent when the speaking style mismatch induced by whispering is present, while the accuracy in the non-mismatched case is near 100 per cent. Similar drop in performance is naturally present in the speaker verification studies where results are reported as equal error rates (EERs).

We found that the comparison of previous studies and their methods is difficult as they have differences in corpora, recognition task (ID/verification), evaluation metric, and in the general approach to address the problem. Some studies focus on reducing the speaking style mismatch by developing features capturing properties that are present in the two speaking styles (*i.e.* enrollment vs. testing), while other studies develop complementary features to be fused in order to tackle the mismatch. Another approach is to transform features of normal speech to resemble those in whispered speech [35], and to use the transformed features in the enrollment phase [35, 36].

As is apparent from the spectrograms in Figure 1, there are clear differences between normal and whispered speech. The differences are especially evident at low frequencies (< 1.5kHz) due to the lack of the periodic voiced excitation in whisper [15]. This property of whispered speech has inspired many previous feature extraction technologies. Some of the features have been extracted from a *limited bandwidth* that exclude lower frequencies altogether. Also, because lower frequencies do not have as important role for whispered speech as for normal speech, features with different frequency warping strategies, such as *linear frequency cepstral coefficients* (LFCCs) and *exponential frequency cepstral coefficients* (EFCCs), have been adopted [9].

Table 1: Overview of previous studies on automatic whispered speech speaker identification or verification. Performances are given for cases: 1) speakers are enrolled and tested with normal speech (n-n) and 2) speakers are enrolled with normal speech but tested using whispered speech (n-w). Even though all the numbers are not exact, they can be used to get a general idea of how much n-w mismatch affects the performance. The number of speakers (#Speakers) tells how many speakers were enrolled in the recognition experiments.

Ref.	**Year**	**Corpus**	**#Speakers (f/m)**	**Methods**	**Backend**	**Performance (n-n / n-w)**	**Metric**
[37]	2007	Described in [37]	22 (n/a)	Feature warping, GMM score combination	GMM	- / -	-
[38]	2008	CHAINS [39]	36 (16/20)	Pykfec features	GMM	80 – 95 % / 15 – 40%	Acc.
[40]	2008	UT-VE I [41]	10 (0/10)	Frequency warping, GMM score competition, LP power spectrum, unvoiced consonant detection	GMM	92% / 53–80%	Acc.
[42]	2009	UT-VE I	10 (0/10)	Limited band LFCCs, feature mapping, LP power spectrum, unvoiced consonant detection	GMM	94% / 48 – 68%	Acc.
[43]	2009	UT-VE I	10 (0/10)	Modified temporal patterns	GMM	- / 44.1–70.4%	Acc.
[9]	2011	UT-VE II [41]	28 (28/0)	LFCC, EFCC, unvoiced consonant detection	GMM	99.2% / 79.3 – 88.4%	Acc.
[44, 35]	2011, 2013	UT-VE I & II	28 (28/0)	Feature transformation from neutral to whisper ([35] extends [45])	GMM-UBM	99.1% / 79.3% – 88.9%	Acc.
[46]	2013	Described in [46]	25 (n/a)	TESBCC, TTESBCC, WIF	GMM	99% / 56%	Acc.
[47]	2013	CHAINS	36 (16/20)	WIF	GMM-UBM	- / -	-
[48]	2015	CHAINS	36 (16/20)	frequency & feature warping, LFCC, WIF, limited band features	GMM-UBM	about 2% / about 30%	EER
[49]	2015	CHAINS, wTIMIT [50]	60 (n/a)	WIF	GMM-UBM, i-vector / PLDA	1.6 – 4.4% / 25.8 – 29.2%	EER
[36]	2016	CHAINS, wTIMIT	60 (n/a)	WIF, feature mapping and fusion	i-vector / PLDA	2.9% / 28.0%	EER
[51]	2017	CHAINS, wTIMIT	60 (n/a)	AAMF, Residual MFCC, limited band features, fusion	i-vector / PLDA	0.9% / 17.8 – 27.3%	EER
[52]	2017	CHAINS	36 (16/20)	EMD-based feature	GMM	8.75 – 9.18% / 13.8 – 14.81%	EER

Being the default feature extraction scheme in speaker recognition, *mel-frequency cepstral coefficients* (MFCCs) [53] have been included in every study listed in Table 1. Therefore, to avoid repetition, we have excluded MFCCs from the listed methods. Some of the features studied were not proposed originally for whispered speech, but as they have shown good performances in other tasks, they have later been adopted to whispered speech in many experiments. These features include *weighted instantaneous frequencies* (WIFs) [47], *pyknogram frequency estimate coefficients* (pykfecs) [38] and *temporal energy subband cepstral coefficients* (TESBCCs, TTESBCCs) [46].

Some of the methods presented in Table 1 harvest long-term properties of speech to the features but this is done in different ways. In [43], subband specific features are extracted with a technique called *modified temporal patterns* (m-TRAPs). More precisely, features are extracted from the horizontal strides of a spectrogram obtained by filtering the spectra with 13 linear filters. In contrast, in the *auditory-inspired amplitude modulation feature* (AAMF) extraction scheme [51], features are extracted from blocks of spectrograms consisting of multiple consecutive short-time frames. AAMFs characterize the rate of change in long-term subband envelopes. As this leads to high-dimensional feature representations, feature selection and principal component analysis (PCA) have been adopted. Feature selection is also used to select features that share the highest amount of mutual information across different speaking styles. The third previous method adopting contextual information extracts features known as the *mean Hilbert envelope coefficients* (MHECs) [54]. These features are closest to the 2DAR based features studied in the present investigation as the MHEC extraction includes smoothing of subband Hilbert envelopes and, similarly to 2DAR, it finally outputs features that resemble standard short-term features.

Recently, *empirical mode decomposition* (EMD) based features have been investigated to extract complementary speech information [52]. Features are extracted from *intrinsic mode functions* (IMFs) and they have shown to boost whispered speaker recognition performance when combined with MFCCs.

3. Properties of whispered speech

The lack of voiced excitation (i.e. periodic glottal flow) in whispered speech [55] is the main aspect that makes whispered speech different from normal speech. The lack of voicing (and thereby also the lack of fundemental frequency and its harmonics) in whispered speech results in reduction of sound energy at low frequencies, which in turn increases spectral flatness. The lack of voicing together with low intensity of the sound makes whispered speech less intelligible than normal speech. Therefore, speakers tend to adapt their voice production mechanisms in other ways to enhance speech intelligibility. These adaptations can be carried out, for example, by changing the vocal tract configuration (affecting formant center frequencies and their bandwidths), speaking rate, or phone durations.

In addition to the lack of voiced excitation, a number of other acoustic differences between normal and whispered speech have been reported. For instance, frequencies of the lowest three formants (F1–F3) tend to be higher in whispered speech [15, 56, 57] with the largest increase in F1. In [15], two other observations were made. First, whispered speech sounds were found to have less energy in frequencies below 1.5 kHz. Second, and rather expectedly, by comparing the average cepstra of individual phone segments, it was shown in [15] that the cepstral distance between voiced utterances of normal speaking style and the corresponding sounds in whispers is greater than the distance between unvoiced sounds in normal speaking style and the corresponding phones in whispers.

In this study, we analyze first how formant (center) frequencies and formant bandwidths differ between whispered and normal speech. Differently from [15, 56, 57] where formants were analyzed either from recordings of isolated vowels or from automatically segmented speech sounds relying on manually segmented training data, we automatically align whispered sentences to their normally spoken counterparts without requiring any speech transcription or manual annotation of segments. After the alignment, the aligned frames are compared to measure differences in formants between whispered and normal speech. Our method can be used not only for isolated vowel utterances but also in processing of realistic, continuous speech, and neither requires it performing manual or automatic speech sound segmentation.

3.1. Corpus description

To analyze formants via aligning normal and whispered speech, a parallel corpus containing utterances spoken in both speaking styles is needed. To this end, we adopt the CHAINS (*CHAracterizing INdividual Speakers*) corpus [39], used especially in recent speaker recognition studies involving whispered speech. The CHAINS corpus is, importantly, also publicly available[1]. The corpus is targeted for advancing the study of speaker identification by investigating unique characteristics of speakers and it contains recordings from 16 females and 20 males speaking in various styles, including normal and whispered speech. Majority of the speakers share the same dialect, spoken in the Eastern part of Ireland. For the formant analysis we included 12 speakers from both genders from the described dialect region. We utilized 33 utterances available in the corpus for each speaker for normal and whispered speaking styles. All the normally spoken samples originate from a single recording session and all the whispered recordings from another session. These sessions were held about two months apart. Speech is sampled at 44.1 kHz, and this sample rate is also used in the formant analysis.

3.2. Analysis of changes in formants via speech alignment

We used VoiceSauce [58] with Praat back-end [59] (Burg's algorithm) to extract formant (center) frequencies and the corresponding formant bandwidths for the lowest three formants for all the utterances. Formants were extracted using 20 ms frame every 2 ms. To make formant tracks less noisy, both formant frequency and bandwidth tracks were median filtered using a 9-frame window.

After extracting the formant data, we considered all pairs of whispered and normal speech where the same speaker spoke the same sentence. Since we have 33 sentence pairs from 12 speakers for both genders (except for one file missing from the original corpus), the total number of sentence pairs is $33 \cdot 12 \cdot 2 - 1 = 791$. We aligned pairs of whispered and normal sentences by using *dynamic time warping* (DTW) [60]. It is apparent from Figure 2 that aligned sentences contain parts where either the alignment can be imprecise or the formant tracks are not reliably estimated. Therefore, we use an automatic detection of reliable segments containing no alignment or formant tracking errors. For details of DTW and the automatic detection of reliable segments, see Appendix A. In panels 2 and 3 of the figure, the segments that are detected to be well aligned are marked with yellow bars. These segments provide aligned formant frequency and bandwidth pairs to be used in analyzing differences in formants between normal and whispered speech.

[1] http://chains.ucd.ie/ (URL accessed March 12, 2018).

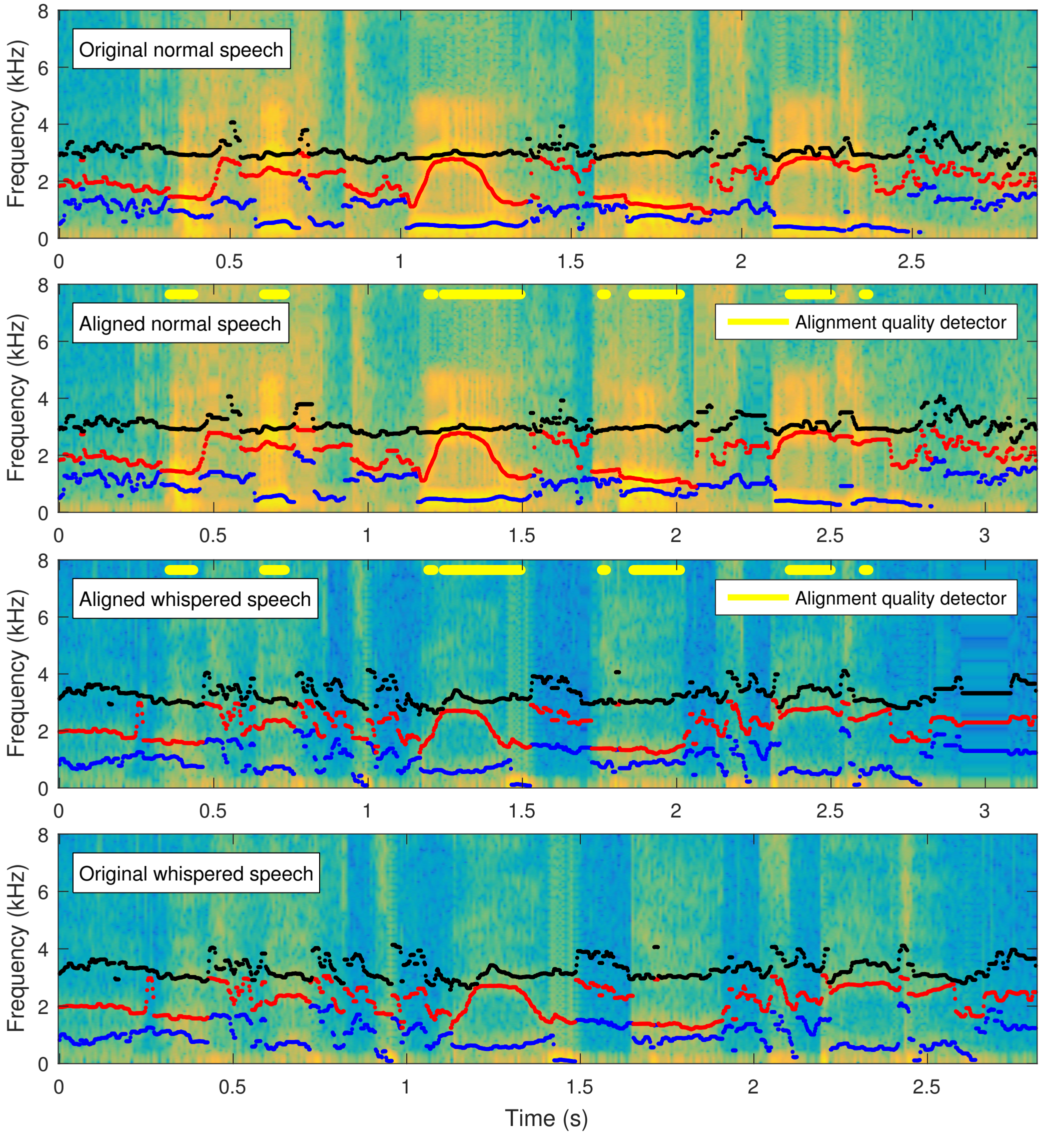


Figure 2: Alignment of normal and whispered speech using dynamic time warping. The first and the last panel of the figure display the spectrograms and formant tracks of the original (non-aligned) pair of normal and whispered sentences spoken by the same speaker. The second and third panel show these sentences after DTW alignment. The aligned sentences are of the same duration, which is longer than the durations of the original sentences because DTW has repeated certain frames multiple times in the aligned speech. After the alignment, an automatic alignment quality detection is applied to the aligned sentences to discard sections of speech where the alignment is unreliable due to, for example, noisy formant tracks or low energy content. We retain the aligned and detected high-quality segments for subsequent analyses.

3.3. Analysis results

We pooled aligned frame pairs and the corresponding aligned formant frequencies and their bandwidths over all speakers and sentences for both genders. Then, we computed histograms of the center frequencies (F1-F3) of the lowest three formants and their bandwidths (B1-B3) for both speaking styles and genders using a 20-Hz bin size. The histograms of the formant frequencies and bandwidths are depicted in Figures 3 and 4, respectively. Further, Table 2 summarizes the mean statistics.

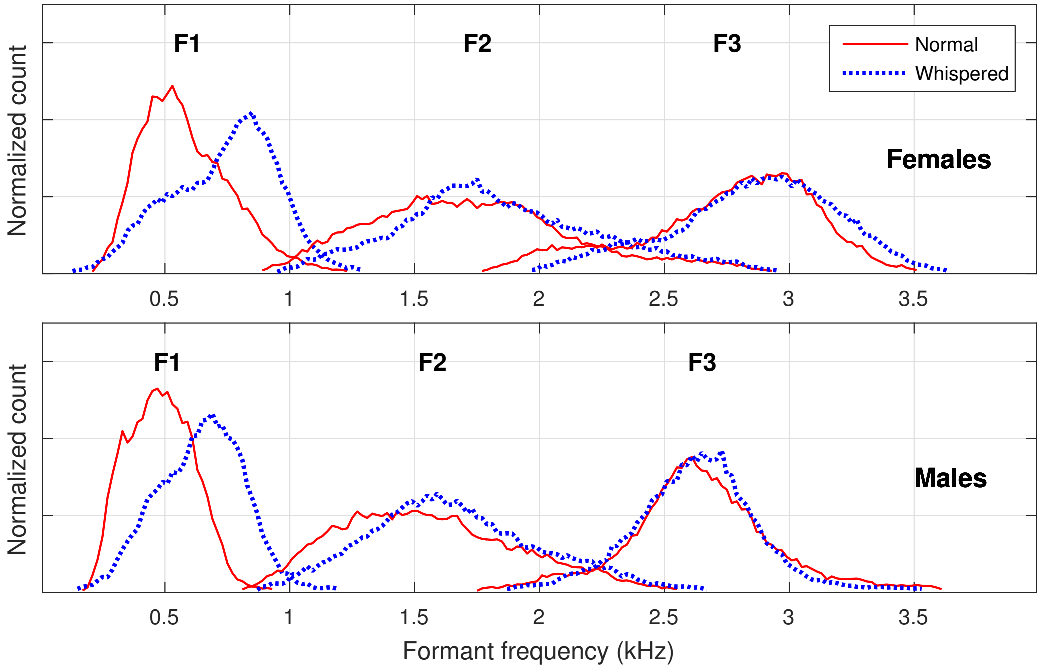


Figure 3: Formant frequency histograms of the lowest three formants (F1–F3) estimated from normal and whispered speech of female and male talkers. The histograms are computed using a 20-Hz bin size. F1 shows the largest change between the two speaking styles.

Distributions in Figure 3 show that formant frequencies tend to be higher in whispered speech. Differences between normal and whispered speech are more prominent for F1 and less so for F2 and F3. On average, for both genders, F1 is about 150 Hz higher and F2 and F3 about 100 Hz higher in whispered speech. An exception is F3 of male speakers, where there is little difference between the two speaking styles. By in large, these observations

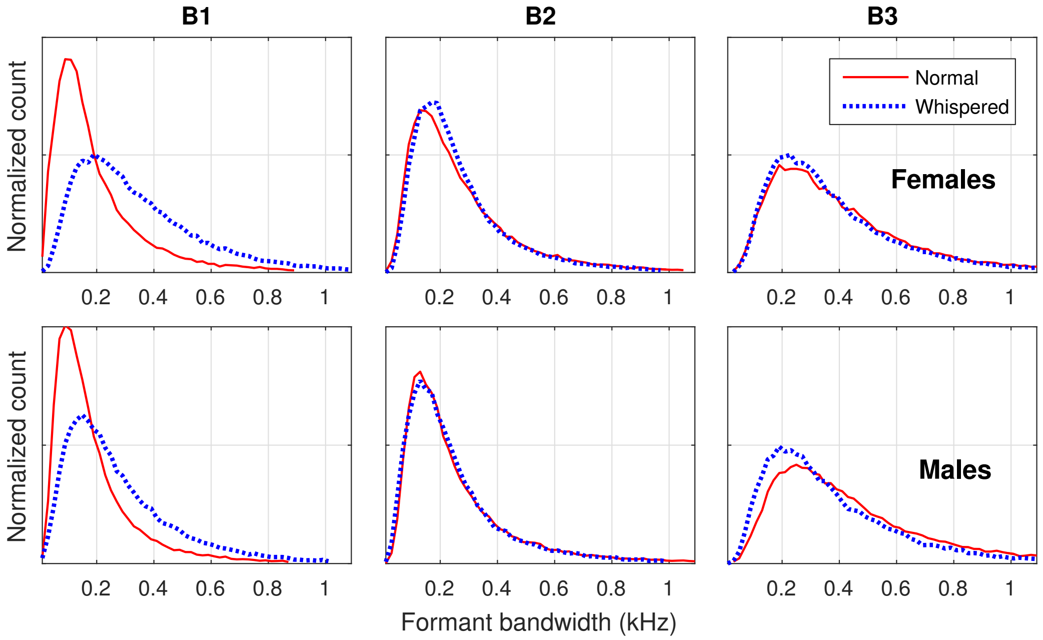


Figure 4: Formant bandwidth histograms of the lowest three formants (B1–B3) estimated from normal and whispered speech of female and male talkers. B1 shows the largest change between the two speaking styles

are in line with the earlier results obtained using different analysis methods [15, 56, 57] for other corpora.

The analysis of formant bandwidths shows that whispered speech tends to have higher B1 whereas B3 tends be higher in normal speech. Bandwidth B2 is similar in both speaking styles.

Table 2: Mean formant frequencies (F1–F3) and bandwidths (B1–B3) in Hz. The standard error of the mean for all values is about 1 Hz.

		F1	**F2**	**F3**	**B1**	**B2**	**B3**
Females	Whispered	746	1853	2861	353	277	396
	Normal	595	1743	2753	206	290	436
	Difference	151	110	108	147	-13	-43
Males	Whispered	665	1675	2642	286	259	382
	Normal	509	1572	2652	195	276	480
	Difference	156	103	-10	91	-17	-98

4. Features for whispered speech speaker recognition

4.1. Two-dimensional autoregressive features

Two-dimensional autoregressive speech modeling (2DAR), introduced in [61], provides a way to construct speech spectrograms that are smoothed both in time and frequency dimensions. As a result of such smoothing, 2DAR spectrograms can be used to extract features that are robust against noise and reverberation without losing relevant information that is used for recognition purposes [7]. The smoothing is first applied in the temporal domain using *frequency domain linear prediction* (FDLP) [29, 30], after which spectral smoothing is done using *time domain linear prediction* (TDLP) [62]. In the following, a brief description of both of these techniques is provided by starting from TDLP, more commonly known as LP (*linear prediction*).

4.1.1. Linear prediction

In conventional LP analysis [62], the current speech sample $x[n]$ is predicted as a weighted sum of the past p samples given by

$$\hat{x}[n] = -\sum_{k=1}^{p} a_k x[n-k] \tag{1}$$

where a_k, $k = 1, \ldots, p$, are known as the *predictor coefficients*. The most common way of solving the predictor coefficients is to minimize the prediction error in the least squares sense. That is, we minimize

$$E = \sum_{n} e^2[n] \tag{2}$$

where $e[n] = x[n] - \hat{x}[n]$. The minimum of (2) is found by calculating partial derivatives with respect to all predictor coefficients a_k and equating them to zero. As a result, we obtain a set of linear equations

$$\sum_{k=1}^{p} a_k r_{ki} = -r_{0i}, \qquad i = 1, \ldots, p, \tag{3}$$

where r_{ki} denotes the *correlation coefficients* given by

$$r_{ki} = \sum_{n} x[n-k]x[n-i].$$

In 2DAR, the *autocorrelation method* [62] is used to solve the predictor coefficients from (3). After solving the coefficients, an *all-pole* estimate of the magnitude spectrum can be obtained as a frequency response of the filter

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}},$$

where G is the gain coefficient.

4.1.2. Frequency domain linear prediction

The first part of 2DAR-modeling, FDLP [29, 30], can be regarded as the frequency domain counterpart of LP. It was shown in [63] that by applying LP to a signal computed by discrete cosine transform (DCT) provides an all-pole estimate of the squared Hilbert envelope of the original time domain signal. In 2DAR, however, FDLP is not applied to the full-band speech signal, but instead to individual frequency bands obtained by windowing the DCT-transformed signal. As a result, all-pole models of the Hilbert envelopes are obtained for each frequency band, and these all-pole models can be used to approximate frequency band energies at regular time instants (*e.g.* once in 10 ms).

4.1.3. Two-dimensional autoregressive modeling

In classical speech analysis, LP is applied by predicting time domain samples within short frames of speech. The 2DAR model, in contrast, models longer-term properties of speech. It achieves this by first reversing the time and frequency domains. This leads to obtaining temporal all-pole power estimates for long-term subbands instead of all-pole spectrum estimates for short-time frames.

The processing steps of 2DAR are depicted in Figure 5. The first step in 2DAR is to transform speech into the frequency domain with DCT. Then, the DCT signal is windowed into subbands using rectangular windows. In this study, we use 100 bands with an overlap of 60% between adjacent bands. These bands are then subjected to the FDLP modeling (LP in frequency domain) to obtain models of the Hilbert envelopes in each band. These envelopes are, in turn, windowed using 25 ms Hamming windows with 60% overlap. Samples of each windowed envelope are integrated to obtain power estimates for each 25 ms time interval of the corresponding frequency band. By stacking power estimates over different subbands, we obtain power spectral estimates for all time-frames. As the next step, the power spectral esti-

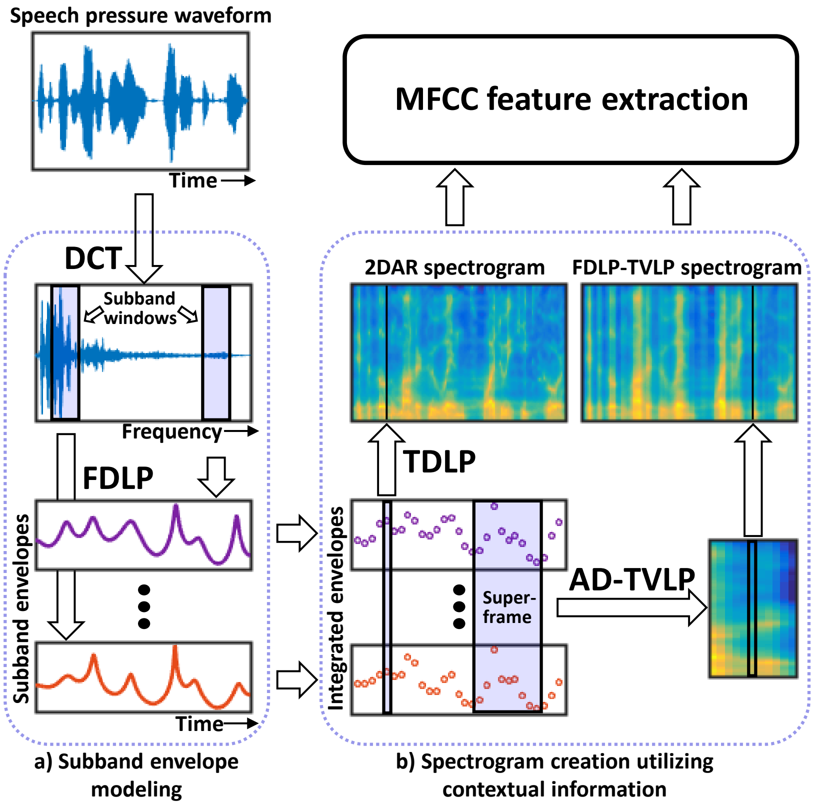


Figure 5: Process flows in creating 2DAR and FDLP-TVLP spectrograms from sentence "If it doesn't matter who wins, why do we keep score?". FDLP applied to subband windows provides time domain envelopes for the subbands. The integrated envelopes shown in the bottom are obtained by summing up values of subband envelopes over 25 ms long Hamming windows that have a 15-ms overlap. Integrated envelopes provide power spectral estimates that are transformed to autocorrelation values and used either in TDLP or autocorrelation domain TVLP (AD-TVLP) modeling. TDLP processing is performed for individual frames, while in AD-TVLP, modeling is performed in superframes that consist of multiple consecutive short-time frames. As the superframe slides forward one frame at a time, only the center spectrum resulting from AD-TVLP modeling of the superframe is retained to the FDLP-TVLP spectrogram.

mates are inverse Fourier-transformed to compute the autocorrelation function. The autocorrelation functions are used in the TDLP, which outputs the final spectral estimates that are used in 2DAR spectrograms and in feature extraction.

4.2. Two-dimensional time-varying autoregressive features

In our preliminary work [1], we have proposed a modification to the 2DAR method that uses *time-varying* linear prediction in the place of conventional LP. In the present study, we cover this technique in a more elaborate manner in both theoretical and experimental means.

4.2.1. Classical time-varying linear prediction

The conventional LP analysis [62] assumes the underlying vocal tract model of a speech signal to remain constant over each short-time interval (frame) of speech. Depending on the frame increment, the model can have abrupt changes from frame to frame. In reality, however, the vocal tract is a continuously varying system that changes even within a single 25-ms frame. TVLP model [31, 32] takes into account the non-stationarity of the vocal tract by allowing the predictor coefficients a_k to be time-varying. Thus, in the case of TVLP (1) becomes

$$\hat{x}[n] = -\sum_{k=1}^{p} a_k[n]x[n-k]. \tag{4}$$

By itself, (4) does not prevent the occurrence of models that change rapidly in time because no constraint has yet been imposed on the change of the predictor coefficients. In TVLP, the rate of the change is constrained by representing the time trajectories of the predictor coefficients as a linear combination of $q+1$ basis functions $\{u_i[n]\}_{i=0}^{q}$ as follows:

$$a_k[n] = \sum_{i=0}^{q} b_{ki}u_i[n]. \tag{5}$$

Typically, basis functions are selected so that they provide smooth, low-pass type of predictor coefficient trajectories. A high number of such basis functions allows for more rapid changes in the predictor coefficients and in the vocal tract model. Conversely, using only one constant basis function, $u_0[n] = 1$, makes the model equivalent to LP. An example of using simple monomial basis function in TVLP modeling of a 50 ms speech frame is given in Figure 6.

In TVLP, minimization of (2) with respect to each basis coefficient b_{ki} leads to a set of equations given by

$$\sum_{k=1}^{p}\sum_{i=0}^{q} b_{ki}c_{ij}[k,l] = -c_{0j}[0,l] \tag{6}$$

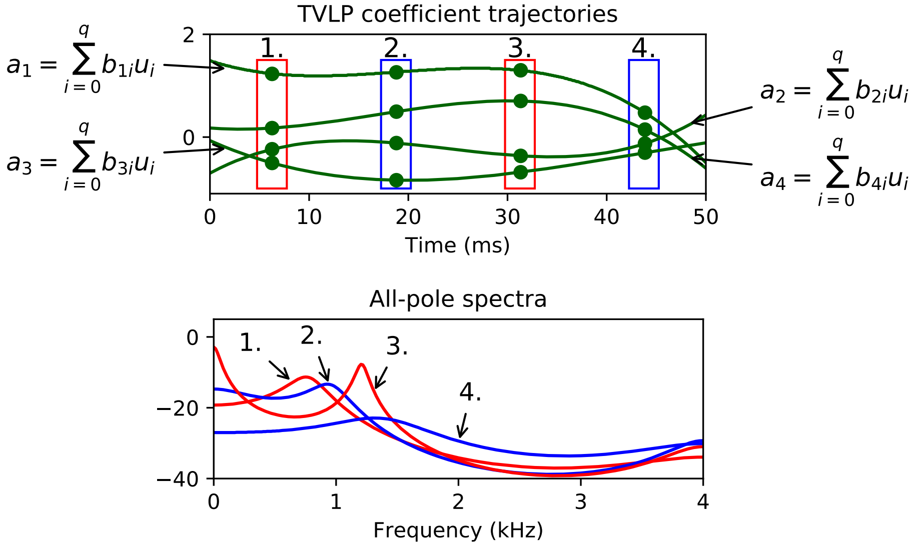


Figure 6: An example of time-varying linear predictive (TVLP) modeling. In this illustration, TVLP coefficient trajectories of a 50 ms long speech frame are modeled using four ($q = 3$) basis functions $u_i, i = 0, \ldots, q$. The model order of TVLP is set to $p = 4$ resulting in four trajectories shown in the upper graph. Because the coefficients are time-varying within the frame, an unique set of predictor coefficients can be sampled at any time instant. The lower graph shows examples of all-pole spectra obtained from four different time instants.

for $1 \leq l \leq p$ and $0 \leq j \leq q$ [31]. Here $c_{ij}[k, l]$ denotes the *generalized* correlation coefficients defined as

$$c_{ij}[k, l] = \sum_{n} u_i[n]u_j[n]x[n-k]x[n-l]. \tag{7}$$

4.2.2. Proposed autocorrelation domain time-varying linear prediction

In classical TVLP formulation, as can be seen from Eq. (7), the computation is directly based on the time-domain signal. However, this can be a serious constraint if one wants to combine the advantages of TVLP with some other robust signal processing techniques. In several robust feature extraction techniques, such as FDLP, *non-negative matrix factorization* (NMF) based feature enhancement [64], and missing data imputation, the signal is converted into a spectro-temporal representation before processing or enhancing it further. In such a scenario, the use of classical TVLP requires the spectro-temporal representations to be converted back to time domain samples. This in turn would require a careful handling of the phase information. In order to avoid distortions that may occur due to phase reconstruction, we propose a modified TVLP analysis which can performed *directly* on the spectro-temporal representations.

Any given spectro-temporal representation $X(t, f)$ can be converted into a sequence of autocorrelation functions by computing the inverse Fourier transform of the power spectrum $X(t, f)$ at each time instant t, given by

$$r_\tau(t) = \frac{1}{2\pi} \int_f X(t, f) \exp(j2\pi f\tau) df \tag{8}$$

At this stage, one can use the conventional LP or TDLP independently on the correlation functions at each time instant by solving the normal equations similar to that in Eq. (3). In such a scenario, the LP coefficients derived at each time instant are prone to errors induced by non-stationary noise and do not take advantage of the fact that the speech production apparatus is a slowly varying inertial system. In order to take advantage of this inertial property of the speech production system, we propose a new TVLP formulation that operates directly on the autocorrelation sequences. This is achieved by imposing a time-continuity constraint on the LP coefficients derived at each time instant by modifying the normal equations in Eq. (3). The resulting normal equations with the continuity constraint, making use of the sequence of autocorrelation functions, is given by

$$\sum_{k=1}^{p} a_k[n] r_{ki}[n] = -r_{0i}[n], \qquad \begin{array}{l} i = 1, \ldots, p, \\ n = 0, \ldots, N-1, \end{array} \tag{9}$$

where $r_{ki}[n]$, $a_k[n]$, and N denote the time-varying autocorrelation coefficients, the time-varying LP coefficients, and the window length for the time-varying analysis, respectively. This expression is similar to Eq. (3), except that both the filter coefficients as well as the correlation coefficients are now functions of time.

Now approximating the piecewise constant filter coefficients $a_k[n]$ using Eq. (5), the above expression in Eq. (9) can be written as

$$\sum_{k=1}^{p} \sum_{j=0}^{q} b_{kj} u_j[n] r_{ki}[n] = -r_{0i}[n], \qquad \begin{array}{l} i = 1, \ldots, p, \\ n = 0, \ldots, N-1, \end{array} \tag{10}$$

This expression is similar to Eq. (6), except that the autocorrelation coefficients $r_{ki}[n]$ do not include basis functions $u_j[n]$ in their computation, as is the case with $c_{ij}[k,l]$ in Eq. (7). This expression can also be interpreted as modeling a piecewise constant filter coefficients $\{a_k[n]; k = 1 \dots p, n = 0 \dots N-1\}$ using a smooth continuous time-varying model represented by $\{b_{ki}; k = 1 \dots p, i = 0 \dots q\}$.

The above set of linear equations can be written in matrix form, given by

$$\boldsymbol{R}\boldsymbol{b} = -\boldsymbol{r} \tag{11}$$

where

$$\boldsymbol{r} = [r_{01}[0], \dots, r_{0p}[0], \dots, r_{01}[N-1], \dots, r_{0p}[N-1]]^T_{Np \times 1} \tag{12}$$

$$\boldsymbol{b} = [b_{10}, \dots, b_{1q}, \dots, b_{p0}, \dots, b_{pq}]^T_{p(q+1)\times 1} \tag{13}$$

$$\boldsymbol{R} = [R_0, R_1, \dots, R_{N-1}]^T_{Np \times p(q+1)}. \tag{14}$$

Here R_n is a $p(q+1) \times p$ matrix whose i^{th} column is given by

$$R_{ni} = \boldsymbol{r}_i[n] \otimes \boldsymbol{u}[n], \tag{15}$$

and $\otimes$ denotes the Kronecker product of $\boldsymbol{r}_i[n] = [r_{1i}[n] \dots r_{pi}[n]]^T$ and $\boldsymbol{u}[n] = [u_0[n] \dots u_q[n]]^T$, given by

$$R_{ni} = [r_{1i}[n]u_0[n], \dots, r_{1i}[n]u_q[n], \dots, r_{pi}[n]u_0[n], \dots, r_{pi}[n]u_q[n]]^T_{p(q+1)\times 1}. \tag{16}$$

The least square solution to the set of linear equations in Eq. (11) can be computed as

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}}{\operatorname{argmin}} \, ||\boldsymbol{r} + \boldsymbol{R}\boldsymbol{b}||_2^2. \tag{17}$$

The above TVLP formulation starting with a sequence of autocorrelation functions is refered to as *autocorrelation domain time-varying linear prediction* (AD-TVLP).

4.2.3. Proposed feature extraction method

Figure 5 illustrates the difference between 2DAR and the proposed TVLP-enhanced version of 2DAR that we call as FDLP-TVLP. After FDLP processing, 2DAR models individual frames with LP, while in our method, we form “superframes” consisting of multiple consecutive frames and feed them to autocorrelation domain TVLP. While LP processing smooths the spectrogram only in the spectral dimension, TVLP is computed *simultaneously* in spectral and temporal domains making the final spectrogram more rubust to noise.

The use of TVLP in the proposed AD-TVLP differs from previous TVLP studies (e.g. [32, 31]) because speech is not modeled in sample-level precision but in frame-precision. While the conventional TVLP formulation uses only the sample-precision in time domain, AD-TVLP can be applied either for sample-precision subband envelopes or for frame-precision integrated envelopes. In our early experiments [1], we found that operating on the frame-level provides better results in SID tasks. Therefore the current study focuses on the frame-level application of AD-TVLP.

4.3. Reference features

We compare 2DAR and and FDLP-TVLP features against multiple reference features. As a first reference feature, we use standard mel-frequency cepstral coefficients (MFCCs) [53] with delta and double-delta coefficients appended. We provide spectral estimates for MFCCs with discrete Fourier transform (DFT). In this work, all the studied features differ only in the spectral estimation part. That is, for every feature, including 2DAR and FDLP-TVLP, we use the same MFCC feature extraction configuration except for the spectral estimation part.

Next, we evaluate features using conventional, short-term LP spectral estimation, which is a justified baseline for the long-term 2DAR and FDLP-TVLP features. We also include *power-law adjusted* LP [8] (LP-α) features, which showed positive results for shouted speech SID in a recent study [8]. The LP-α is a spectral compression technique to reduce the effect of the spectral tilt difference between normal and shouted speech. Since the spectral tilt varies also between normal and whispered speech, it was justified to select also (LP-α) as one robust reference feature extraction method in the current study. Then, the FDLP method without TDLP or TVLP is included as a reference method containing temporal processing but without the spectral processing. Finally, as a last reference spectral estimation method, we use *minimum variance distortionless response* (MVDR) [33] spectrum.

5. Experimental set-up for speaker recognition

5.1. Speech corpus and experimental protocols

To conduct speaker recognition experiments from whispered speech, a suitable evaluation corpus needs to be identified first. Unlike studying speech of normal speaking style, for which a large supply of corpora and associated standard evaluation protocols are available, there are fewer databases available for studying speaker recognition from whispered speech. With the help of data given in Table 1, we decided to adopt the CHAINS corpus [39] in the current study. In the recognition experiments, the original speech data, sampled at 44.1 kHz, were downsampled to 16 kHz.

To cover a broad set of possible application scenarios, we designed two speaker recognition evaluation protocols. The first one, the speaker identification (SID) protocol, is relevant in applications such as personalized control of smart devices. The second one, the automatic speaker verification (ASV), is relevant in applications such as user authentication for access control, forensics, and surveillance. While there are several prior studies on speaker recognition from whispered speech (Table 1), there exists, unfortunately, no commonly used standard protocol. Hence, we decided to design our own protocols with an intention of maximizing the number of recognition trials with a limited amount of data.

5.1.1. Speaker identification protocol

The SID protocol utilizes recordings from 12 females and 12 males from the same dialect region (Eastern Ireland). For each speaker, we utilized 32 spoken sentences available in the corpus for normal and whispered speaking styles. The original corpus, in fact, contains 33 utterances, but we excluded one of them since one audio file was missing from the original corpus distribution.

Similar to [8], we adopt a *leave-one-out* protocol to increase the number SID evaluation trials: we leave one utterance at a time, to be used as the test trial, and use the remaining 31 utterances to train the target speaker model. As the average duration of an utterance is 2.81 seconds, the average duration of speech data to train the speaker model is about 87 seconds.

One SID trial consists of comparing the test utterance against all the 12 speaker models of the same gender. The identified speaker is the one whose target speaker model reaches the highest SID score. This way, we have in total $12 \times 32 = 384$ SID trials per gender. We conducted SID trials in two ways. First, by using normal speech for both speaker enrollment and testing and second, by using normal speech in enrollment but whispered speech in testing.

As our objective measure of performance, we compute speaker identification rate, defined as the proportion of correctly classified test segments to the total number of scored test segments, computed separately per each gender.

5.1.2. Speaker verification protocol

The ASV protocol utilizes all of the normal and whispered speech data in the CHAINS corpus in order to obtain more trials and to increase the reliability of the results. That is, we use all 33 sentences and 4 fables (typically 30 − 60 seconds long) from 36 speakers (16 females, 20 males). The first fable, with an average duration of 56 seconds, is used for training the target speaker models. The remaining 3 fables are cut into 3 second long clips and they are used together with the 33 sentences as test segments in the verification experiments. By testing all test segments against all speaker models, we obtain trial statistics summarized in Table 3. Again, trials were conducted in two ways by always enrolling speakers with normal speech, but testing either with normal or whispered speech. The designed protocol is similar to the one in [48], with a difference that our test set is somewhat larger.

Table 3: Number of trials in the speaker verification protocol. The numbers of same-speaker trials are given in parentheses.

	Normal	Whispered
Females	16,752 (1,047)	17,136 (1,071)
Males	25,520 (1,276)	26,000 (1,300)
All	42,272 (2,323)	43,136 (2,371)

We report verification performances in terms of equal error rate (EER), the rate at which false alarm and miss rates are equal. When comparing the proposed features to the reference features, we also report 95 % confidence intervals of EERs, computed using the methodology of [65]. That is, confidence interval around EER is EER ± c, where

$$c = 1.96\sqrt{\frac{\text{EER}(1-\text{EER})}{4N_i} + \frac{\text{EER}(1-\text{EER})}{4N_s}},$$

where N_i is the number of impostor trials and N_s is the number of same speaker trials.

5.2. Speaker recognition system

We performed speaker recognition experiments with a classic *Gaussian mixture model – universal background model* (GMM-UBM) system [19]. While there are many other possible choices for the back-end, including i-vectors, the GMM-UBM tends to provide comparative (or higher) accuracy for short utterances [66, 67, 68] and is suitable with limited development datasets, requiring only UBM training data specification besides the enrollment and test samples. For each of the feature extraction techniques, we train a 256-component UBM using the TIMIT corpus, which is sampled at 16 kHz and recorded in quiet environments. To make the UBM training data gender-balanced, we use 192 speakers for both genders. The target speaker models are obtained by *maximum a posteriori* (MAP) adaptation [69] of the UBM using the training sentences of a particular speaker. A relevance factor of 2 was used to adapt the Gaussian component means of the UBM.

5.3. Feature configurations

The feature extraction techniques compared in this study differ substantially in their internal computations. At the output, however, they all yield estimates of the power spectrum (or power-spectrum like presentation) that are computed in 25-ms frames, incremented in 10-ms steps. For the power spectrum estimation, we study the following five reference methods besides the proposed FDLP-TVLP method: discrete Fourier transform (DFT), linear prediction (LP) [62], power-law adjusted LP (LP-α) [8], frequency domain linear prediction (FDLP) [70], minimum variance distortionless response (MVDR) [33], and 2-dimensional autoregressive model (2DAR) [7]. The power spectrum, estimated using one of these methods, is used as input to the MFCC computation chain in the standard way [71]. In the identification experiments, the center frequency of the first and last mel-filter were set to 200 Hz and 7800 Hz, respectively, whereas in verification experiments, we adopt a narrower frequency range between 200 Hz and 5600 Hz (In Section 6.4, we study how the feature extraction bandwidth affects the system performance.) We use 19 MFCCs without the energy coefficient, appended with delta and double delta coefficients, yielding 57-dimensional feature vectors. MFCCs are RASTA-filtered [72] except when temporal processing with

FDLP is used, as it had a negative effect on the SID performance. Including both RASTA and FDLP could cause too much temporal smoothing of speech information. For the other spectrum estimation methods, RASTA had a positive or neutral effect. Finally, MFCCs of non-speech frames are discarded and the remaining MFCCs are normalized to have zero mean and unit variance per utterance.

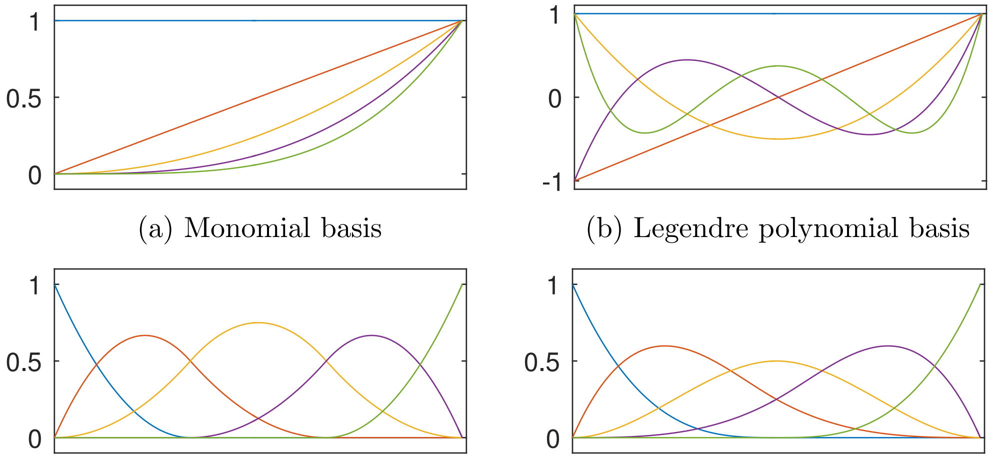
Each of the feature extraction techniques have a number of control parameters that need to be set. For LP, we found the model order of at least 40 to yield the highest SID accuracy. Thus, in this study, we use $p = 40$ for LP, LP-α, and MVDR. We use the same model order for TDLP in 2DAR and for AD-TVLP in FDLP-TVLP. For both 2DAR and FDLP-TVLP, we found that a FDLP model order of 24 or higher for one second long segments provides the best performance for both normal and whispered speech. Because of the varying utterance lengths, we normalize the FDLP prediction order according to the length of the processed utterance. In the identification protocol, we use an FDLP model order of $p = 24$ and for the verification, we set the model order to $p = 48$. For LP-α, the best α value was found to be 0.05.

6. Speaker recognition results

In this section, we provide results of the conducted speaker recognition experiments. First, we optimize the control parameters of the proposed FDLP-TVLP feature extraction method and then continue by comparing the method to the reference methods. We provide results for two kinds of speaker recognition tasks, speaker identification (SID) and speaker verification. Further, we address the SID task in more detail by analyzing SID accuracies at the level of individual speakers.

6.1. The choice of basis functions for time-varying linear prediction

In TVLP, temporal contours of LP filter coefficients are modeled as a linear combination of basis functions. Many types of functions, such as *Monomial functions* [73], *trigonometric functions* [31], and *Legendre polynomials* [74], have been used previously. The choice of basis functions, however, has not been studied for the AD-TVLP formulation used in this study where we model the LP predictor coefficient trajectories at frame precision instead of



(a) Monomial basis

(b) Legendre polynomial basis

(c) 3rd order B-spline basis

(d) 4th order B-spline basis

Figure 7: Different types of basis functions used in AD-TVLP.

sample-by-sample basis as is done in the classic TVLP. Therefore, we analyze the impact of the choice with four kinds of basis functions illustrated in Figure 7.

First, we study the effect of superframe size (N) and the number of used basis functions ($q+1$) together with a monomial basis (Figure 7a)

$$u_i[n] = n^i, \quad i = 0, \ldots, q, \quad n = 0, \ldots, N-1, \tag{18}$$

used in our previous work [1]. The superframe size determines the number of adjacent frames being fed to the AD-TVLP model at once. The results presented in Table 4 indicate that the parameter choice is not critical, provided that the superframe size is sufficiently large and that the number of basis functions is large enough for a given superframe size. A suitable number of basis functions seems to be around one-third of the superframe size.

Table 4: Effect of the superframe size and the number of basis functions to the speaker verification performance (EER (%)) using monomial basis.

	Number of basis functions ($q+1$)							
Superframe size	Normal vs. normal				Normal vs. whisper			
# frames (ms)	3	4	5	6	3	4	5	6
7 (85 ms)	3.4 ± 0.4	3.4 ± 0.4	3.3 ± 0.4	3.5 ± 0.4	27.9 ± 0.9	28.3 ± 0.9	28.4 ± 0.9	28.9 ± 0.9
11 (125 ms)	3.7 ± 0.4	3.3 ± 0.4	**3.1** ± 0.4	3.5 ± 0.4	28.0 ± 0.9	**27.4** ± 0.9	28.1 ± 0.9	28.1 ± 0.9
15 (165 ms)	5.0 ± 0.5	4.5 ± 0.4	3.5 ± 0.4	3.5 ± 0.4	29.2 ± 0.9	29.0 ± 0.9	28.0 ± 0.9	28.7 ± 0.9
19 (205 ms)	6.0 ± 0.5	5.1 ± 0.5	4.2 ± 0.4	3.7 ± 0.4	30.4 ± 1.0	29.2 ± 0.9	28.8 ± 0.9	28.4 ± 0.9
23 (245 ms)	7.7 ± 0.6	6.5 ± 0.5	5.2 ± 0.5	4.6 ± 0.4	32.4 ± 1.0	30.5 ± 1.0	29.6 ± 0.9	28.8 ± 0.9

Table 5: Speaker verification equal error rates (%) for different basis types (superframe size = 11).

	Number of basis functions							
	Normal vs. normal				Normal vs. whisper			
Basis type	3	4	5	6	3	4	5	6
Monomial	3.7 ± 0.4	3.3 ± 0.4	**3.1** ± 0.4	3.5 ± 0.4	28.0 ± 0.9	**27.4** ± 0.9	28.1 ± 0.9	28.1 ± 0.9
Legendre	3.7 ± 0.4	3.3 ± 0.4	**3.1** ± 0.4	3.5 ± 0.4	28.0 ± 0.9	**27.4** ± 0.9	28.1 ± 0.9	28.1 ± 0.9
3rd order B-spline	8.3 ± 0.6	3.5 ± 0.4	3.4 ± 0.4	3.3 ± 0.4	32.6 ± 1.0	28.1 ± 0.9	27.8 ± 0.9	28.6 ± 0.9
4th order B-spline	–*	6.8 ± 0.5	3.3 ± 0.4	3.2 ± 0.4	–*	31.7 ± 1.0	28.2 ± 0.9	28.5 ± 0.9

* undefined configuration

We fix the superframe size to $N = 11$ for the remaining experiments with other basis function types, namely Legendre polynomials (Figure 7b) and B-splines (Figure 7c, 7d) [75]. In contrast to monomials and Legendre polynomials, B-splines have a local support. However, the results in Table 5 indicate that this does not provide benefits to the given ASV task. Further, the monomial and Legendre bases provide equal results. Therefore, we consider only the monomial basis with 4 basis functions for all the remaining experiments.

6.2. Model orders for spectral and temporal processing of speech

As the 2DAR scheme performs linear prediction in both frequency (FDLP) and time (TDLP) domains, it has two main parameters to be optimized. The model order for FDLP determines the amount of smoothing in temporal subband envelopes; lower value resulting in more smoothed spectrograms in time. TDLP model order, in turn, is used to control the amount of details present in the frequency dimension. In [7], the effect of model orders of 2DAR to speaker verification performance was studied for clean and noisy speech.

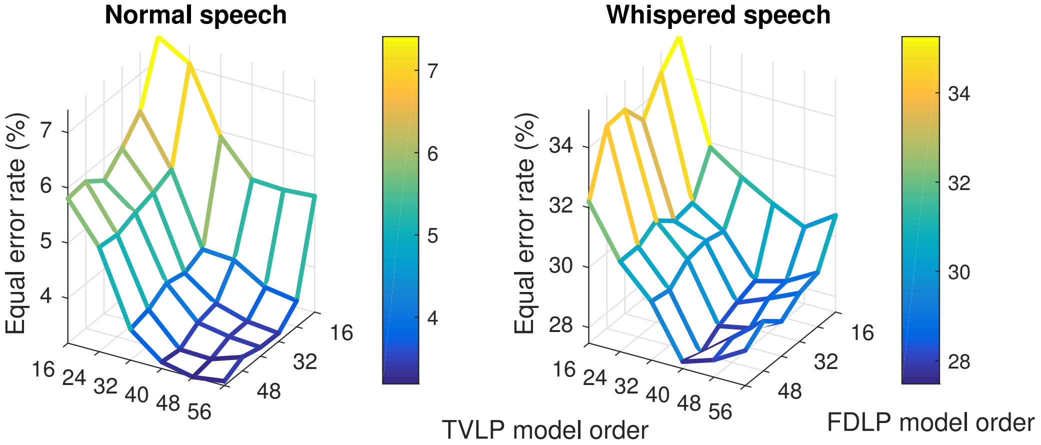


Figure 8: Speaker verification equal error rates (%) for normal and whispered speech using different model orders in FDLP-TVLP modeling.

Similarly, the proposed FDLP-TVLP contains two model order parameters, one for FDLP and the other for AD-TVLP. In Figure 8, we jointly vary both in the speaker verification task. Regarding to the FDLP model order, our results are very similar to the ones in [7]. In specific, the model order has to be at least 24 to obtain good results.

Concerning the model order in spectral processing, there are some differences largely explained by the differences in data. In [7], the experiments performed with TDLP revealed that a high model order (> 20) is better for clean speech while for a noisy speech, a low model order (< 15) improves the performance. As our data is clean and has twice as high sampling rate (16 kHz), we find that for the AD-TVLP and for 16 kHz sampling rate, the best performance is obtained with model orders higher than 32.

6.3. Comparison of features in speaker identification task

Table 6 presents the identification results for all the evaluated spectrum estimation methods described in Sections 4 and 5.3. All the methods provide close to or equal to 100.0 % identification rate when there is no speaking style mismatch present between enrollment and testing. With whispering-induced mismatch, however, all the accuracies drop to around 50 %. Unlike in many other speaker recognition studies (*e.g.* [8], [71]), we curiously find that female speakers obtain higher identification accuracies than males for whispered speech. From the whispered SID studies listed in Table 1, only [48] reports SID results per gender basis and did not find considerable differences in performances between males and females on the same corpus and same type of back-end. The difference might be partly explained by differing sampling rate (16 vs. 8 kHz) and the evaluation protocol. In addition, unlike in the current study, gender dependent UBMs were used in [48].

Table 6: SID accuracies (%) for different spectrum estimation methods.

	Normal vs. normal			Normal vs. whisper		
Method	Females	Males	All	Females	Males	All
DFT	100.0	100.0	100.0	52.1	40.4	46.2
LP	100.0	100.0	100.0	51.3	44.8	48.0
LP-α ($\alpha = 0.05$)	100.0	100.0	100.0	53.9	42.4	48.2
MVDR	100.0	100.0	100.0	51.0	35.4	43.2
FDLP	96.1	94.8	95.4	39.3	32.6	35.9
2DAR	99.7	99.5	99.6	55.2	44.0	49.6
FDLP-TVLP	99.7	99.7	99.7	**56.5**	**45.3**	**50.9**
FDLP-TVLP-α	99.7	99.5	99.6	54.9	41.4	48.2

We have grouped the methods in Table 6 into three categories. The first group consists of the two standard short-term methods, DFT and LP, from which LP provides higher SID accuracy. Then, in the second group, the LP-α and MVDR methods have been introduced to provide added robustness to short-term features. From these two, only LP-α outperforms or matches the DFT and LP baselines when subjected to whispered speech SID. The last group consists of the FDLP-derived methods that use long-term speech processing. The FDLP method, by itself, is behind most of the short-term methods but improves substantially when combined with the spectral processing provided by LP (2DAR).

Finally, the proposed FDLP-TVLP method has a moderate margin to 2DAR and provides the best overall performance for whispered test cases. We also tried to include the α-compression of the power spectrum to the FDLP-TVLP method prior to the TVLP processing step, but as the results show, we did not find this to be beneficial. This might be due to both methods already having similar beneficial effects by themselves, achieved through different means. An aggressive α-compression can be used to make spurious spectral peaks less prominent, but similar effect can be achieved using contextual information, as in FDLP-TVLP, by smoothing the spectra over time.

The obtained SID results, as a whole, imply more benefits being gained by improving spectral processing as opposed to temporal processing. This is supported by the good results obtained with LP and LP-α and by the large performance difference between FDLP and the other FDLP-based methods that include LP-based spectral processing. On the other hand, the proposed FDLP-TVLP achieved the best performance by including two layers of temporal processing, one by FDLP, and the other by AD-TVLP. This suggests that TVLP methodology, in the context of style mismatch compensation, is worthwhile of further studies.

6.4. The choice of frequency range in feature extraction

Next, we studied how the frequency range used in the MFCC extraction affects the identification results of DFT and FDLP-TVLP for whispered speech. We kept the first mel-filter centered at 200 Hz and changed the position of the other filters according to the position of the last mel-filter, which was varied between 4000 Hz and 7600 Hz. The results are presented in Figure 9. We find that as the frequency range decreases the identification accuracy drops. Furthermore, we find that FDLP-TVLP does not seem to benefit from the inclusion of higher frequencies ($> 5000 Hz$) as much as DFT.

In prior studies [48, 42], the frequency range has been limited by increasing the frequency of the first mel-filter. It has been found that discarding spectral information below 1 kHz improves system performance in normal-whispered mismatched test cases, since the spectral differences between the two speaking modes are largest in the low frequency range.

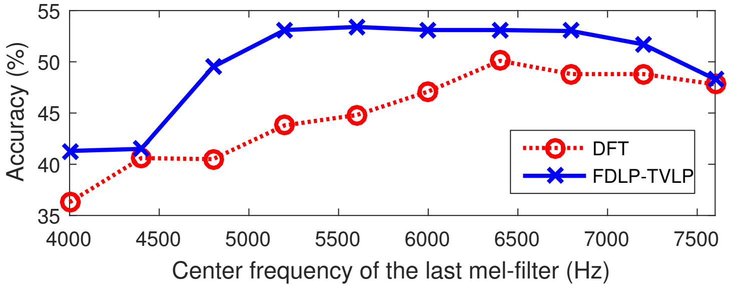


Figure 9: Speaker identification accuracies (%) for whispered speech using different frequency ranges in the MFCC computation.

6.5. Speaker-by-speaker analysis

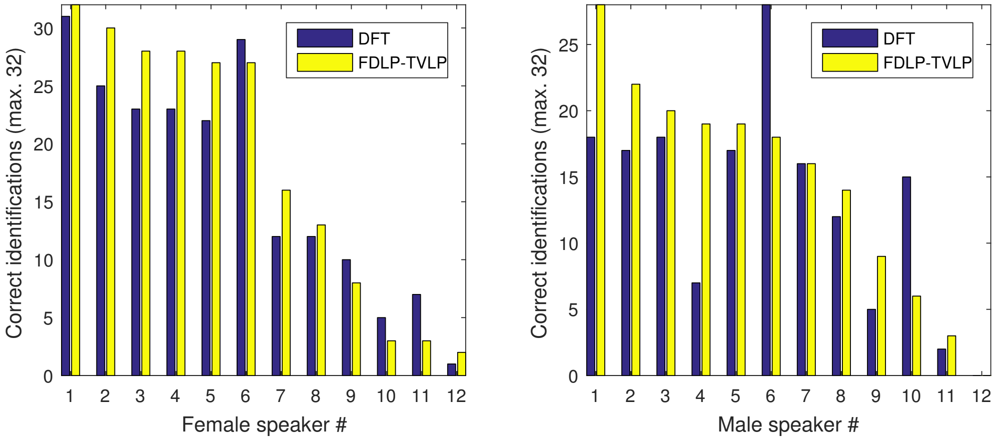


Figure 10: Number of correct identifications for each individual speaker in normal-whisper mismatched case.

Up to this point, we have shown the results in a pooled form over all the speakers. With an aim to provide further insights into SID from whispered speech, we analyze results on a speaker-by-speaker basis. Figure 10 displays the SID results of DFT and FDLP-TVLP methods for each speaker, sorted

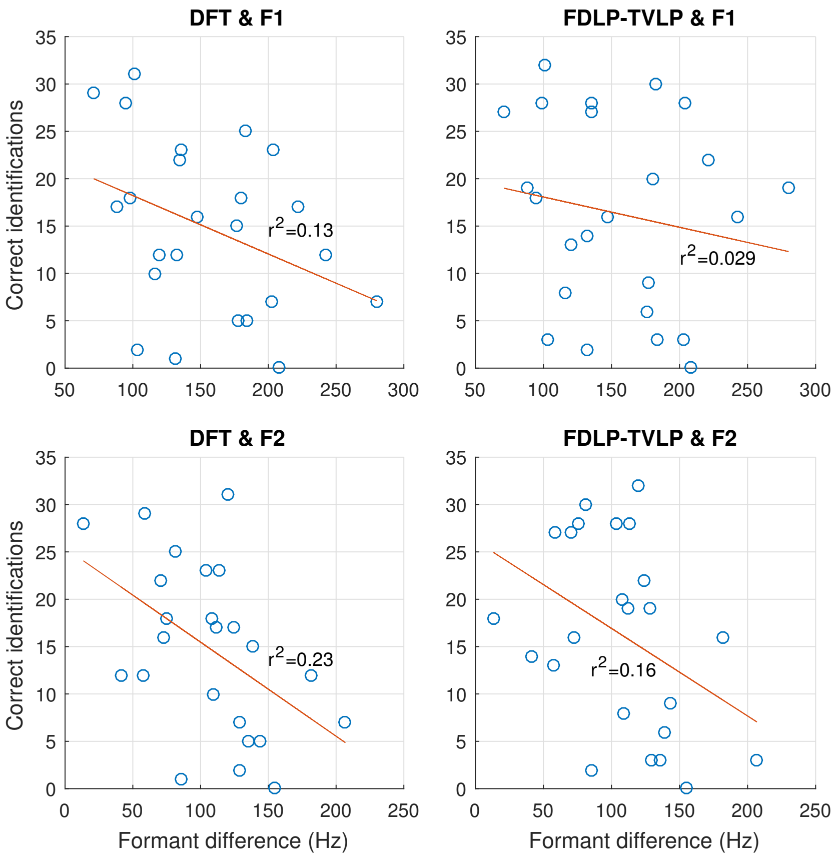


Figure 11: The number of correct identifications for a speaker versus mean difference of formants frequencies of whispered and normal speech. Results are shown for F1 (top) and F2 (bottom) and for feature extraction methods DFT (left) and FDLP-TVLP (right). If the formant differences are large for a speaker, the number of correct identifications is more likely to be small.

according to the number of correct identifications obtained using FDLP-TVLP. The results indicate large differences between individuals; some speakers are correctly identified almost every time, while others are almost always misidentified. From informal listening of the most difficult speakers, we could not identify any obvious abnormal speech characteristics or recording quality related issues. Hence, instead, we decided to analyze whether a change in formant values between normal and whispered speech explains the differences between SID performances of individual speakers. In Figure 11, we present correlations between the number of correct identifications for a speaker and the average difference of formant frequencies between whispered and normal speech. We find a weak correlation between these two variables. Interestingly, the correlation is stronger for F2 than for F1 and it is also stronger for DFT than for FDLP-TVLP, which suggests that FDLP-TVLP might tolerate formant changes slightly better.

6.6. Comparison of features in speaker verification task

Table 7: Speaker verification equal error rates (%) with 95% confidence intervals [65] for different spectrum estimation methods.

	Normal vs. normal		
Method	Females	Males	All
DFT	2.67 ± 0.50	**2.50** ± 0.44	**2.57** ± 0.33
LP	**2.48** ± 0.49	3.11 ± 0.49	2.84 ± 0.35
LP-α ($\alpha = 0.05$)	3.15 ± 0.55	3.37 ± 0.51	3.45 ± 0.38
MVDR	3.56 ± 0.58	3.61 ± 0.52	3.61 ± 0.39
FDLP	3.82 ± 0.60	4.58 ± 0.59	4.30 ± 0.42
2DAR	3.34 ± 0.56	2.97 ± 0.48	3.10 ± 0.36
FDLP-TVLP	3.06 ± 0.54	3.33 ± 0.50	3.27 ± 0.37
FDLP-TVLP-α	3.78 ± 0.60	3.91 ± 0.55	3.86 ± 0.40

	Normal vs. whisper		
Method	Females	Males	All
DFT	26.24 ± 1.36	29.89 ± 1.28	29.69 ± 0.95
LP	26.32 ± 1.36	30.38 ± 1.28	28.38 ± 0.93
LP-α ($\alpha = 0.05$)	26.12 ± 1.36	30.38 ± 1.28	28.91 ± 0.94
MVDR	24.90 ± 1.34	31.45 ± 1.29	28.13 ± 0.93
FDLP	26.70 ± 1.37	31.69 ± 1.30	30.33 ± 0.95
2DAR	25.38 ± 1.35	29.69 ± 1.27	28.43 ± 0.93
FDLP-TVLP	**24.84** ± 1.34	**29.08** ± 1.27	**27.48** ± 0.92
FDLP-TVLP-α	25.96 ± 1.36	30.32 ± 1.28	28.64 ± 0.94

The results for the speaker verification task are presented in Table 7. As expected, the results resemble those obtained from the identification experiments. For normally spoken speech, male and female performances are close to each other. For whispered speech, however, there is a clear gap between genders; in specific, females show 3–6 % lower EERs in absolute terms. As before, DFT and LP show the best performance in the normal speaking style, while for whispered speech, FDLP-TVLP gives the lowest error rates although it compromises performance in normal speech by about 0.5% (absolute EER). As a general finding, all the methods yield high error rates when tested under speaking style mismatch. Our results are similar to those reported in [48].

6.7. Analysis of local variability in spectrogram estimation methods

In machine learning, the well-known over-fitting vs. under-fitting trade-off, also know as the *bias-variance* trade-off, relates to generalization of models beyond a given training set. Models with a comparatively larger number of degrees of freedom tend to produce good results on training data (low bias) but fail to generalize (high variance). Being constrained by the low-order polynomial functions, the TVLP models addressed in this study are intuitively more rigid in comparison to the traditional way of extracting MFCCs. For this reason, we expect them to be less sensitive to acoustic mismatch between enrollment and test utterances, including changes in speaking style.

As our last analysis, we are interested to objectively quantify the degree of feature rigidness directly from the spectral representations. To this end, inspired by the widespread use of *Laplace operator* in image processing [76], and by viewing spectrograms as images, we adopt discrete Laplacians to quantify the rigidness of spectrogram representation of speech signals both in time and frequency variables obtained by different spectrum estimators. In specific, we use the average value of absolute values of Laplacian evaluated at all points of spectrograms (excluding non-speech segments). The discrete Laplacian $\mathcal{L}$ is defined as,

$$\mathcal{L}(t,f) = S(t-1,f) + S(t+1,f) + S(t,f-1) + S(t,f+1) - 4S(t,f),$$

where t and f refer to indices of time and frequency values, respectively, and where $S(t,f)$ is a speech spectrogram. Furthermore, to measure variability along one dimension only, we similarly use,

$$\begin{aligned}\mathcal{L}_t(t,f) &= S(t-1,f) + S(t+1,f) - 2S(t,f) \quad \text{and}\\ \mathcal{L}_f(t,f) &= S(t,f-1) + S(t,f+1) - 2S(t,f).\end{aligned}$$

We computed the spectrograms of the UBM data (TIMIT) using DFT, LP, and FDLP-TVLP methods. The average absolute Laplacians are presented in Table 8. In comparison to DFT, we find that LP helps to reduce the local variability in time (mean($|\mathcal{L}_f|$)) as it smooths spectra in frequency.

As a side product, it also reduces variability in time (mean($|\mathcal{L}_t|$)) because the noisy values of spectrogram get removed. The smoothing in time in FDLP-TVLP causes a large drop to variability in time while the variability in frequency is similar to the LP method.

Table 8: Analysis of local variability in spectrograms obtained using DFT, LP, and FDLP-TVLP spectrum estimators. Variabilities are measured as average absolute values of Laplacians extracted from speech spectrograms. Laplacian $\mathcal{L}$ is used the measure variability jointly in both dimensions and $\mathcal{L}_t$ and $\mathcal{L}_f$ are used to measure variability independently in time an frequency, respectively. As the means are computed over a large dataset, standard errors of the means in all cases are less than 0.01, making all the values significantly different from each other.

	mean($|\mathcal{L}|$)	mean($|\mathcal{L}_t|$)	mean($|\mathcal{L}_f|$)
DFT	33.36	18.52	18.76
LP	13.88	9.49	5.20
FDLP-TVLP	7.09	2.90	4.78

7. Conclusions

Significant advancements on speaker recognition research have been made in recent years by speaker modeling using i-vector and DNN technology, yet mismatch conditions due to the intrinsic and extrinsic variabilities remain as a major cause of performance degradation. In the current study, we addressed the problem of mismatch arising from a specific speaking style, whispering. Besides providing an up-to-date and self-contained tutorial survey on speaker recognition from whispered speech, we introduced a new speech modeling technique that involves a long-term speech analysis based on a joint utilization of *frequency domain linear prediction* and *time-varying linear prediction* (FDLP-TVLP).

Our speaker recognition experiments on the CHAINS corpus indicate that speaker recognition from whispered speech can benefit from using FDLP-TVLP when the control parameters of the model are properly set. We made the following conclusions regarding the parameter choices. The number of basis functions for the proposed TVLP method should be about one third of the number of short-time frames in the superframe. With an experiment using four basis functions, we conclude that recognition performances do not depend much on the type of the basis function. We recommend to use simple monomial bases. Further, we experimented with different model orders for

FDLP and TVLP, and we found that as long as the model orders are above 24 and 32 (assuming 16 kHz sampling rate) respectively, performance remains high.

In comparison with baseline reference features, we have found that the FDLP-TVLP feature performs considerably better than standard MFCC and LP-based MFCC features for speaker recognition from whispered speech. On the other hand, we observe a small performance degradation with normal voice. Also, the results obtained with the proposed feature have shown considerable improvement over closely related FDLP and 2DAR feature, and these indicate that speech modeling including the *time-varying* form of linear prediction helps for the recognition of whispering speaker. Interestingly, the recognition performance for whispered female voice is better than for the male voice. This finding contradicts with the usual observation in speaker recognition experiments where recognition of female speakers is more difficult than male speakers.

From speaker-by-speaker analysis of speaker identification performance, we observed considerable accuracy differences across the speakers. This suggests that the articulatory process for producing whispered voice is highly dependent on the individual person and evidently, some speakers are naturally good at disguising themselves by producing close to unidentifiable whispered voice. From the more detailed analysis, we found that a small part of individual differences can be explained by the amount of changes in formant frequencies between the normal and whispered speaking styles.

While our preliminary study on whispered speech showed promising results, we are aware of the following limitations planned to be addressed in future studies. Firstly, although the current study shows moderate improvement over baseline, the identification accuracy for *normal vs. whisper* condition is almost half of the accuracy obtained for the non-mismatched *normal vs. normal* condition. One reason for this is the absence of whispered data in the back-end training where we used the TIMIT data, which is well suited only for *normal vs. normal* condition. Secondly, for processing whispered speech, we have applied exactly same processing steps as used for the normal speech. One advantage of this approach is that it does not use knowledge of the underlying speaking style during processing, however, at the cost of a possible performance loss when compared to speaking style specific processing used jointly with a speaking style detector. Thirdly, the current study uses GMM-UBM framework which does not explicitly consider any channel or session variability compensation technique as used in i-vector-PLDA frameworks. For such advanced systems, preparing suitable data recipe for training parameters and hyper-parameters is difficult due to the lack of appropriate and adequate data. Finally, the study was conducted on a relatively small dataset requiring further experiments to be conducted to generalize the existing results.

We found that comparison of findings with the existing studies on speaker recognition from whispered speech is difficult due to the lack of commonly used data sets and evaluation protocols. In addition to having standard evaluation protocols, the research community would benefit from a large publicly available corpus containing recordings of both normal and whispered speech. A larger corpus would allow the use of more data-intensive methods and would make the evaluation of research findings more reliable.

Appendix A. Details on aligning normal and whispered speech

Appendix A.1. Frame-to-frame distance function

To perform time alignment of normal and whispered speech with dynamic time warping (DTW), we defined a frame-to-frame distance function d given by

$$d(\text{frame}_i, \text{frame}_j) = |\text{F1}_i - \text{F1}_j| + |\text{F2}_i - \text{F2}_j| + |\text{F3}_i - \text{F3}_j| + |E_i - E_j| + 500,$$

where F1–F3 are the formant (center) frequencies in Hz and E is the log energy of a frame computed between 4 kHz and 8 kHz . That is, the distances are based on computing absolute differences of the formant frequencies and the log energy values. The reason for excluding low frequencies (0–4 kHz) from the energy computation is that, due to lack of the fundamental frequency, the energy in whispered speech differs more from that of normal speech in the low frequency range than in the high frequency range. Before distance computation, the log energies are shifted and scaled so that for each sentence the minimum log energy is 0 and the maximum log energy is 1000. In addition, we add a constant term 500 to all of the distances to reduce the amount of time stretching in the DTW algorithm.

Appendix A.2. Automatic alignment quality detection

The detection of well aligned segments consists of three steps. First, energy based speech activity detection is performed for both normal and whispered speech to discard non-speech frames. Second, we discard those segments whose formant tracks can be considered unreliable. The algorithm for detecting formant tracking quality uses 30-frame long sliding window to discard windows that contain too many sudden jumps (more than 200Hz) between the consecutive formant frequencies. Finally, we discard segments that contain considerable amount of time stretching (repeated frames). More precisely (within a 30-frame window), if the sum of the repeated frames in aligned normal and aligned whispered speech is more than 8, the window will be discarded. An example of a segment that contains too much time stretching can be seen near the 3 second mark in the third panel of Figure 2.

References

[1] V. Vestman, D. Gowda, M. Sahidullah, P. Alku, T. Kinnunen, Time-varying autoregressions for speaker verification in reverberant conditions, in: INTERSPEECH, 2017, pp. 1512–1516.

[2] D. A. Reynolds, R. C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE transactions on Speech and Audio Processing 3 (1) (1995) 72–83.

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

[4] S. J. Prince, J. H. Elder, Probabilistic linear discriminant analysis for inferences about identity, in: IEEE 11th International Conference on Computer Vision (ICCV 2007), IEEE, 2007, pp. 1–8.

[5] P. Rajan, A. Afanasyev, V. Hautamäki, T. Kinnunen, From single to multiple enrollment i-vectors: Practical PLDA scoring variants for speaker verification, Digital Signal Processing 31 (2014) 93–101.

[6] G. Liu, Y. Lei, J. H. Hansen, Robust feature front-end for speaker identification, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012), IEEE, 2012, pp. 4233–4236.

[7] S. Ganapathy, S. H. Mallidi, H. Hermansky, Robust feature extraction using modulation filtering of autoregressive models, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 22 (8) (2014) 1285–1295.

[8] R. Saeidi, P. Alku, T. Bäckström, Feature extraction using power-law adjusted linear prediction with application to speaker recognition under severe vocal effort mismatch, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 24 (1) (2016) 42–53.

[9] X. Fan, J. H. Hansen, Speaker identification within whispered speech audio streams, IEEE transactions on audio, speech, and language processing 19 (5) (2011) 1408–1421.

[10] J. H. Hansen, M. K. Nandwana, N. Shokouhi, Analysis of human scream and its impact on text-independent speaker verification, The Journal of the Acoustical Society of America 141 (4) (2017) 2957–2967.

[11] R. G. Hautamäki, M. Sahidullah, V. Hautamäki, T. Kinnunen, Acoustical and perceptual study of voice disguise by age modification in speaker verification, Speech Communication 95 (2017) 1–15.

[12] H. Masthoff, A report on a voice disguise experiment, Forensic Linguistics 3 (1996) 160–167.

[13] H. J. Künzel, Effects of voice disguise on speaking fundamental frequency, International Journal of Speech Language and the Law 7 (2) (2000) 149–179.

[14] J.-C. Junqua, The lombard reflex and its role on human listeners and automatic speech recognizers, The Journal of the Acoustical Society of America 93 (1) (1993) 510–524.

[15] T. Ito, K. Takeda, F. Itakura, Analysis and recognition of whispered speech, Speech Communication 45 (2) (2005) 139–152.

[16] J. H. L. Hansen, T. Hasan, Speaker recognition by machines and humans: A tutorial review, IEEE Signal Process. Mag. 32 (6) (2015) 74–99.

[17] P. Kenny, Joint factor analysis of speaker and session variability: theory and algorithms, technical report CRIM-06/08-14 (2006).

[18] P. Kenny, Bayesian speaker verification with heavy-tailed priors, in: Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, 2010, p. 14.

[19] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker verification using adapted Gaussian mixture models, Digital signal processing 10 (1-3) (2000) 19–41.

[20] L. Li, D. Wang, C. Zhang, T. F. Zheng, Improving short utterance speaker recognition by modeling speech unit classes, IEEE/ACM Transactions on Audio, Speech, and Language Processing 24 (6) (2016) 1129–1139.

[21] H. Zeinali, H. Sameti, L. Burget, HMM-based phrase-independent i-vector extractor for text-dependent speaker verification, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (7) (2017) 1421–1435.

[22] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, K. Yu, Deep feature for text-dependent speaker verification, Speech Communication 73 (2015) 1 – 13.

[23] D. Garcia-Romero, X. Zhou, C. Y. Espy-Wilson, Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012), IEEE, 2012, pp. 4257–4260.

[24] P. Rajan, T. Kinnunen, V. Hautamäki, Effect of multicondition training on i-vector PLDA configurations for speaker recognition., in: INTERSPEECH, 2013, pp. 3694–3697.

[25] G. Heigold, I. Moreno, S. Bengio, N. Shazeer, End-to-end text-dependent speaker verification, in: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, IEEE, 2016, pp. 5115–5119.

[26] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, S. Khudanpur, Deep neural network-based speaker embeddings for end-to-end speaker verification, in: Spoken Language Technology Workshop (SLT), 2016 IEEE, IEEE, 2016, pp. 165–170.

[27] T. Kinnunen, H. Li, An overview of text-independent speaker recognition: From features to supervectors, Speech communication 52 (1) (2010) 12–40.

[28] C. Zhang, J. H. Hansen, Analysis and classification of speech mode: whispered through shouted, in: INTERSPEECH, Vol. 7, 2007, pp. 2289–2292.

[29] J. Herre, J. D. Johnston, Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS), in: Audio Engineering Society Convention 101, Audio Engineering Society, 1996.

[30] R. Kumaresan, A. Rao, Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications, The Journal of the Acoustical Society of America 105 (3) (1999) 1912–1924.

[31] M. G. Hall, A. V. Oppenheim, A. S. Willsky, Time-varying parametric modeling of speech, Signal Processing 5 (3) (1983) 267–285.

[32] D. Rudoy, T. F. Quatieri, P. J. Wolfe, Time-varying autoregressions in speech: Detection theory and applications, IEEE Transactions on audio, Speech, and Language processing 19 (4) (2011) 977–989.

[33] M. N. Murthi, B. D. Rao, All-pole modeling of speech based on the minimum variance distortionless response spectrum, IEEE Transactions on Speech and Audio Processing 8 (3) (2000) 221–239.

[34] P. X. Lee, D. Wee, H. S. Y. Toh, B. P. Lim, N. F. Chen, B. Ma, A whispered mandarin corpus for speech technology applications, in: INTERSPEECH, 2014, pp. 1598–1602.

[35] X. Fan, J. H. Hansen, Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams, Speech communication 55 (1) (2013) 119–134.

[36] M. Sarria-Paja, M. Senoussaoui, D. O'Shaughnessy, T. H. Falk, Feature mapping, score-, and feature-level fusion for improved normal and whispered speech speaker verification, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), IEEE, 2016, pp. 5480–5484.

[37] Q. Jin, S.-C. S. Jou, T. Schultz, Whispering speaker identification, in: IEEE International Conference on Multimedia and Expo, IEEE, 2007, pp. 1027–1030.

[38] M. Grimaldi, F. Cummins, Speaker identification using instantaneous frequencies, IEEE Transactions on Audio, Speech, and Language Processing 16 (6) (2008) 1097–1111.

[39] F. Cummins, M. Grimaldi, T. Leonard, J. Simko, The chains corpus: Characterizing individual speakers, in: Proc of SPECOM, Vol. 6, 2006, pp. 431–435.

[40] X. Fan, J. H. Hansen, Speaker identification for whispered speech based on frequency warping and score competition, in: INTERSPEECH 2008, pp. 1313–1316.

[41] C. Zhang, J. H. Hansen, Advancements in whisper-island detection within normally phonated audio streams, in: INTERSPEECH, 2009, pp. 860–863.

[42] X. Fan, J. H. Hansen, Speaker identification with whispered speech based on modified LFCC parameters and feature mapping, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009), IEEE, 2009, pp. 4553–4556.

[43] X. Fan, J. H. Hansen, Speaker identification for whispered speech using modified temporal patterns and MFCCs, in: INTERSPEECH 2009, pp. 896–899.

[44] X. Fan, J. H. Hansen, Speaker identification for whispered speech using a training feature transformation from neutral to whisper, in: INTERSPEECH, 2011, pp. 2425–2428.

[45] X. Fan, J. H. Hansen, Acoustic analysis for speaker identification of whispered speech, in: IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2010), IEEE, 2010, pp. 5046–5049.

[46] N. P. Jawarkar, R. S. Holambe, T. K. Basu, Speaker identification using whispered speech, in: International Conference on Communication Systems and Network Technologies (CSNT 2013), IEEE, 2013, pp. 778–781.

[47] M. Sarria-Paja, T. H. Falk, D. O'Shaughnessy, Whispered speaker verification and gender detection using weighted instantaneous frequencies, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), IEEE, 2013, pp. 7209–7213.

[48] M. O. Sarria-Paja, T. H. Falk, Strategies to enhance whispered speech speaker verification: A comparative analysis, Canadian Acoustics 43 (4) (2015) 31–45.

[49] M. Sarria-Paja, M. Senoussaoui, T. H. Falk, The effects of whispered speech on state-of-the-art voice based biometrics systems, in: IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE 2015), IEEE, 2015, pp. 1254–1259.

[50] B. P. Lim, Computational differences between whispered and non-whispered speech, Ph.D. thesis, University of Illinois at Urbana-Champaign (2011).

[51] M. Sarria-Paja, T. H. Falk, Fusion of auditory inspired amplitude modulation spectrum and cepstral features for whispered and normal speech speaker verification, Computer Speech & Language 45 (2017) 437–456.

[52] R. Sharma, S. Prasanna, R. Bhukya, R. Das, Analysis of the intrinsic mode functions for speaker information, Speech Communication 91 (2017) 1 – 16.

[53] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE transactions on acoustics, speech, and signal processing 28 (4) (1980) 357–366.

[54] S. O. Sadjadi, J. H. Hansen, Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification, speech communication 72 (2015) 138–148.

[55] V. C. Tartter, Whats in a whisper?, The Journal of the Acoustical Society of America 86 (5) (1989) 1678–1683.

[56] W. F. Heeren, Vocalic correlates of pitch in whispered versus normal speech, The Journal of the Acoustical Society of America 138 (6) (2015) 3800–3810.

[57] M. Higashikawa, K. Nakai, A. Sakakura, H. Takahashi, Perceived pitch of whispered vowels-relationship with formant frequencies: A preliminary study, Journal of Voice 10 (2) (1996) 155–158.

[58] Y.-L. Shue, P. Keating, C. Vicenik, K. Yu, Voicesauce: A program for voice analysis, in: Proceedings of the ICPhS XVII, 2011, pp. 1846–1849.

[59] P. Boersma, Praat: doing phonetics by computer (version 6.0.29, http://www.praat.org/.

[60] D. Ellis, Dynamic time warp (dtw) in matlab, Web resource, available: http://www. ee. columbia. edu/ dpwe/resources/matlab/dtw.

[61] M. Athineos, H. Hermansky, D. Ellis, PLP^2: Autoregressive modeling of auditory-like 2-D spectro-temporal patterns, in: ISCA Tutorial and Research Workshop (ITRW) on Statistical and Perceptual Audio Processing, 2004.

[62] J. Makhoul, Linear prediction: A tutorial review, Proceedings of the IEEE 63 (4) (1975) 561–580.

[63] M. Athineos, D. P. Ellis, Autoregressive modeling of temporal envelopes, IEEE Transactions on Signal Processing 55 (11) (2007) 5237–5245.

[64] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, G. Rigoll, The munich 2011 chime challenge contribution: Nmf-blstm speech enhancement and recognition for reverberated multisource environments, in: Machine Listening in Multisource Environments, 2011.

[65] S. Bengio, J. Mariéthoz, A statistical significance test for person authentication, in: Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop, no. EPFL-CONF-83049, 2004.

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

[67] H. Zeinali, H. Sameti, L. Burget, J. Cernockỳ, N. Maghsoodi, P. Matejka, i-vector/HMM based text-dependent speaker verification system for reddots challenge., in: INTERSPEECH, 2016, pp. 440–444.

[68] S. Dey, P. Motlicek, S. Madikeri, M. Ferras, Template-matching for text-dependent speaker verification, Speech Communication 88 (2017) 96–105.

[69] C.-H. Lee, J.-L. Gauvain, Speaker adaptation based on map estimation of hmm parameters, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-93), 1993, Vol. 2, IEEE, 1993, pp. 558–561.

[70] S. Thomas, S. Ganapathy, H. Hermansky, Recognition of reverberant speech using frequency domain linear prediction, IEEE Signal Processing Letters 15 (2008) 681–684.

[71] J. Pohjalainen, C. Hanilçi, T. Kinnunen, P. Alku, Mixture linear prediction in speaker verification under vocal effort mismatch, IEEE Signal Processing Letters 21 (12) (2014) 1516–1520.

[72] H. Hermansky, N. Morgan, RASTA processing of speech, IEEE transactions on speech and audio processing 2 (4) (1994) 578–589.

[73] L. A. Liporace, Linear estimation of nonstationary signals, The Journal of the Acoustical Society of America 58 (6) (1975) 1288–1295.

[74] Y. Grenier, Time-dependent arma modeling of nonstationary signals, IEEE Transactions on Acoustics, Speech, and Signal Processing 31 (4) (1983) 899–911.

[75] J. Friedman, T. Hastie, R. Tibshirani, The elements of statistical learning, second edition, Springer series in statistics New York, 2009.

[76] X. Wang, Laplacian operator-based edge detectors, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (5) (2007) 886–890.