Teager–Kaiser Energy Operators for Overlapped Speech Detection

Navid Shokouhi, *Student Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—Overlapped speech is referred to a monophonic audio signal in which at least two speakers are present at the same time. In this study, the focus is on distinguishing overlapped from single-speaker speech, i.e., overlapped speech detection. We develop an overlap detection algorithm using an enhanced time-frequency representation, called Pyknogram, estimated directly from the input audio signal. Pyknograms use the Teager–Kaiser energy operator to detect resonant time-frequency units and thereby suppress nonharmonic structures. We show how the resulting Pyknograms provide high separability in terms of detecting the presence of interfering speech. Our proposed unsupervised Pyknogram-based detection results in over 30% relative improvement in overlap detection error rates across different signal-to-interference ratios (SIR) compared to baseline systems. In addition, a case study is presented where we evaluate speaker verification performance under different overlap conditions using the GRID database and observe that speaker verification equal error rates (EER) vary from 2% to 30%, depending on the average SIR values introduced to train and test sets. In order to estimate the reliability of speaker verification scores across different trials, overlap detection results are interpreted as low-level information and *stack*ed alongside verification outputs. The resulting high-dimensional space is passed through a support vector machine classifier to find the separating hyperplane between target and imposter scores. Combining overlap detection scores with speaker verification on average yields 20% relative decrease in EER. We also provide an upper bound for this approach using existing overlap labels, which yields 23% relative improvement.**

***Index Terms*—Co-channel speech, overlap detection, Teager-Kaiser energy operators.**

I. Introduction

OVERLAPPED speech is referred to a monophonic audio signal in which at least two speakers are simultaneously active. Single-channel recordings from meetings or conversations are examples during which speakers may overlap. Separating the resulting mixture becomes especially difficult when one does not assume prior knowledge about speaker identities or speech content. Most studies on overlapped speech have focused on separating the target or suppressing interfering speech [1]. Often to de-noise and thereby improve the performance of automatic speech applications [2]–[4] (primarily speech recognition). However, over the past decade, due to vast developments in recognition systems such as speaker recognition and diarization, a growing trend of detecting overlapped regions has been observed. In speaker recognition, the presence of interfering speech in conversational speech styles not only reduces the effectiveness of trained speaker models but also increases the uncertainty in scoring test files with overlapped regions [5]. Removing overlapped segments increases model reliabilities which consequently improves recognition [6]. State-of-the-art speaker diarization systems are also currently at a stage where one of the main sources of error is the presence of overlapped speech [7], [8]. Overlaps are a source of confusion in speaker diarization systems, since there is no basis for selecting ground-truth in overlapped regions. This makes evaluating speaker diarization systems more challenging. A reasonable work-around is to ignore overlapped regions when evaluating diarization performance. Fortunately, for applications such as speaker recognition and diarization it is rarely necessary to separate the target from interfering speaker in overlapped speech, since preserving speech content is not a priority. One can improve system performance by detecting and excluding overlapped segments for both speaker recognition and diarization. This task, which replaces interferer suppression and target separation with overlapped speech detection, is sometimes called "usable speech detection"[1] [5]. An overlapped speech detection system can be used in any of the aforementioned tasks as a data purification step or a signal processing front-end. Fig. 1 shows how overlap detection is used in three important speech processing applications (speaker diarization, recognition, and speech recognition). In addition to these applications, overlap detection can be used as a conversation analysis tool, since overlaps are highly correlated with conversational cues. The correlation between overlaps and long-term conversational cues can either be used to detect overlaps [9], or it can be used to analyze conversations through overlap detection [10], [11].

Traditionally, studies have used spectral harmonicity as a key component in detecting overlapped speech [12], [13]. This

Manuscript received July 15, 2016; revised November 15, 2016 and January 11, 2017; accepted January 12, 2017. Date of publication March 6, 2017; date of current version April 7, 2017. This work was supported in part by AFRL under Contract FA8750-15-1-0205 and in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tomi Kinnunen.

N. Shokouhi is with the Center for Robust Speech Systems, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: nxs113020@utdallas.edu).

J. H. L. Hansen is with the Center for Robust Speech Systems, Erik Jonsson School of Engineering and Computer Science, Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75080-1407 USA (e-mail: john.hansen@utdallas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2017.2678684

[1]In order to avoid any confusion between this study and the assumptions made in [5], we use the more general term overlapped speech detection.

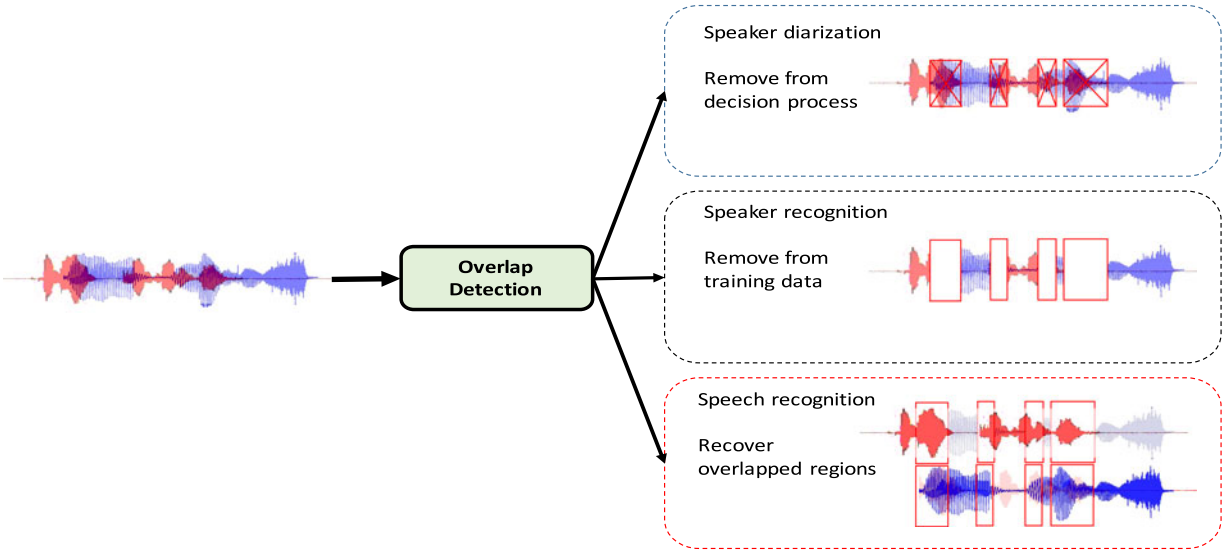

Fig. 1. Applications of overlap detection. Top: In speaker diarization, removing ignoring overlapped regions provides a more fair assessment of diarization performance. Middle: Removing overlaps from in speaker recognition increases the reliability of training data. Bottom: Overlap detection results can be used as an initial step to recover overlapped regions.

approach is motivated by the fact that two fundamental frequencies exist in many instances of overlapped speech which disarranges the harmonic structure observed in single-speaker speech. In [14], the peak-to-valley ratios in frame-based spectral autocorrelations are introduced as a discriminating feature for overlapped speech detection through the same assumption. Spectral flatness measure, the ratio of geometric to arithmetic means calculated from spectral bins in a speech frame, has also been used as a measure to capture harmonicity and has been used to detect the presence of overlapped speech [12]. Another related characteristic is observed when monitoring fundamental frequencies along time. Adjacent pitch period comparison (APPC) presented in [15] uses the temporal variation of estimated "pitch" periods as a measure to detect "usable" speech with the assumption that temporal variations of adjacent pitch periods are significantly higher in overlap. A multi-pitch tracking algorithm proposed in [16] was used in [17] to estimate coexisting fundamental frequencies in the presence of multiple speakers. Regions where more than one fundamental frequency is estimated are labeled as overlap. The multi-pitch tracking technique described in [16], decomposes speech into sub-bands and pitch estimation is only performed on reliable sub-bands.

A slightly different, yet fundamentally similar, approach to distinguish overlapped speech is to use speech kurtosis which measures higher order moments of the signal statistics [18]. The change in kurtosis is a consequence of increased signal complexity in the presence of two speech sources. Alternative measures have been proposed to magnify signal complexity at overlaps by modeling non-linearities [19], [20]. A conclusive summary of common features used to detect overlapped speech for improved speaker diarization is presented in [21]–[23].

A number of studies have considered investigating spectral characteristics at formant frequency locations when dealing with overlapped speech. Giuliani *et al*. use a filter-based approach to improve speech recognition rates for different instances of meeting conditions by adding a detection step that separates double-speaker speech from single-speaker audio [24]. This was accomplished by cascading two-layer sub-band filters to capture formant characteristics. Formant frequency information was obtained by filtering the signal at sub-bands with center frequencies and bandwidths corresponding to nominal F_1, F_2, and F_3 values for all English vowels. One of the reasons Formant-based overlapped speech analysis has received less attention is the difficulties in modeling pole interactions at overlapped regions, which is an issue for linear predictive modeling and other commonly used formant tracking techniques.

In this study, we use the AM-FM speech model along sub-bands [25] to model resonances. An energy operator based approach [25], [26] is used to track harmonics in each sub-band (overlapped or single-speaker) and analyze the signal in those regions to determine whether speech is overlapped. Energy operators have previously been used to deal with signals with more than one source [27] (aka co-channels signals[2]). Maragos *et al*. use higher order energy operators to develop an algorithm that simultaneously demodulates the components of a co-channel mixture in AM-FM modulated signals [27]. Litvina *et al.* separate speech from music using the Teager energy operator (TEO) separation algorithm [25] [28], where they used the extracted components to design a time-varying filter and suppress the interfering signal. Similar multicomponent signal decomposition techniques have been addressed using energy operators to separate narrow-band signals [29]–[31].

Our goal is to incorporate sub-band analysis to design a technique suitable for *overlapped speech detection*. The motivation for sub-band decomposition is to be able to use TEO methods on narrow-band components and detect speech harmonics. The

[2]Co-channel is a more general terminology used to described multi-component signals. In the case of speech, co-channel speech may refer to any single-channel recording that contains speech from multiple speakers, regardless of whether there is overlap.

present study is an extension to [32] that proposed using Pyknograms as overlap detection measures. The focus of [32] was to use overlap detection results to improve word-count estimation in realistic conversational speech data from the Prof-life-log [33] corpus. This study provides a detailed description of overlap detection using Pyknograms.

Section II provides a description of Pyknogram extraction for overlap detection. Pyknogram extraction is split into two steps: 1) frequency estimation, which finds resonance frequencies across the time-frequency spectrum (Section II-A); 2) frequency selection, which prunes out unreliable estimates (Section II-B). Section II also presents our proposed overlap detection measure based on dynamic changes observed in Pyknograms (Section II-C). Section III describes the experiments, by 1) listing the baseline features used for overlap detection (Section III-A), 2) presenting the dataset adopted from the speech separation challenge (Section III-B), and 3) evaluating overlap detection performance in different SIR conditions (Section III-C) and under different segment durations (Section III-D). Sections IV and V investigate the impact of overlapped speech on various aspects of speaker verification. Finally, Section VI summarizes this study and presents an outline to future works.

II. Overlapped Speech Detection

Detecting overlapped segments has previously been considered in tasks such as speaker recognition and diarization [5], [21]. In such problems, the presence of a secondary speaker either decreases model reliability (in training), or introduces confusion in the decision-making process by distorting test files. In cases where speech is of contextual value, such as in speech recognition, the traditional approach is to somehow magnify the presence of a target speaker or weaken interfering speakers. Unfortunately, removing unwanted speech at overlaps is not straightforward and requires prior knowledge of one or both speakers. Such difficulties further motivate the use of overlapped speech detection. Detecting overlaps is computationally advantageous when one has the luxury of neglecting overlapped data [5]; as is the case in speaker recognition and diarization [34]. This study proposes a method for overlap detection in monophonic speech. By detecting overlapped speech segments, we are able to remove them from the training and decision-making process.

We propose a novel approach for overlapped speech detection based on an enhanced spectrogram. These spectrograms, called Pyknograms, were first introduced by Potamianos and Maragos in [35], [36] and are calculated by applying multi-band demodulation in the AM-FM speech model framework [25].[3] Pyknograms provide a more prominent representation of harmonic trajectories, which we propose to use as a means to detect the presence of interfering speech.

[3]The authors in [36] used the term "Pyknogram" which stems from the Greek word "pykno" meaning dense. Pyknograms represent highly resonating regions in time-frequency plots as populated scatter plots, hence the term density.

A. Pyknogram Extraction - Frequency Estimation

In Pyknograms [36], the harmonic structure of speech is enhanced by decomposing spectral sub-bands into amplitude and frequency components. This sub-band analysis uses the AM-FM speech model [25] to decompose speech sub-bands and thereby calculate corresponding instantaneous frequencies and bandwidths. To extract Pyknograms, the speech signal is initially passed through a filter-bank (we have modified the algorithm to use logarithmically spaced Gamma-tone filters, while [36] uses linearly-spaced Gabor filters). Filter-bank outputs ($x_i(n)$, in which i represents filter indexes) are then decomposed into amplitude and frequency components using the discrete energy separation algorithm (DESA-1) [25], where the per sample frequency, $f_i(n)$, and amplitude, $a_i(n)$ (shown in Fig. 2 for a given speech sub-band). Frequency and amplitude estimates for the ith sub-band, $x_i(n)$, are:

$$f_i(n) = \frac{1}{2\pi}\arccos\left(1 - \frac{\Psi[x_i(n) - x_i(n-1)]}{2\Psi[x_i(n)]}\right), \quad (1)$$

$$|a_i(n)| = \sqrt{\frac{\Psi[x_i(n)]}{\sin^2(2\pi f_i(n))}}, \qquad i = 1, 2, ..., N_s \quad (2)$$

where n is the time sample. N_s is the number of sub-bands in the filter-bank and $\Psi(.)$ is the discrete energy operator, defined for any given signal, $x(n)$, as:

$$\Psi[x(n)] = x^2(n) - x(n-1)x(n+1). \quad (3)$$

A weighted average of the instantaneous frequencies, F_w, is estimated over 25 msec windows (aka frames), indexed by t. Together sub-band analysis and time framing results in time-frequency units (t, i), where i corresponds to the frequency sub-band index and t corresponds to frames [37]. Instantaneous frequencies are weighted using the estimated signal power ($|a_i(n)|^2$). The average frequency computed for each time-frequency unit can be viewed as the 1st-order moment of instantaneous frequencies.

$$F_w(t, i) = \frac{\sum_{n_t}^{n_t+T-1} f_i(n) a_i^2(n)}{\sum_{n_t}^{n_t+T-1} a_i^2(n)}, \quad (4)$$

T is the number of samples per frame, from $n = n_t$ to $n = n_t + T - 1$, in which n_t is the beginning sample of frame t. The algorithm also provides a way to estimate weighted bandwidths for the frequency component, (5). What we refer to here as bandwidths are essentially 2 nd-order frequency moments.

$$B_w(t, i) = \sqrt{\frac{\sum_{n_t}^{n_t+T-1} (\dot{a}_i(n)/2\pi)^2 + (f_i(n) - F_w(t, i))^2 a_i^2(n)}{\sum_{n_t}^{n_t+T-1} a_i^2(n)}}, \quad (5)$$

where $f_i(n)$ and $a_i(n)$ are instantaneous frequency and amplitude values from (1) and (2). In (4), the instantaneous frequencies are averaged over the tth frame using squared instantaneous

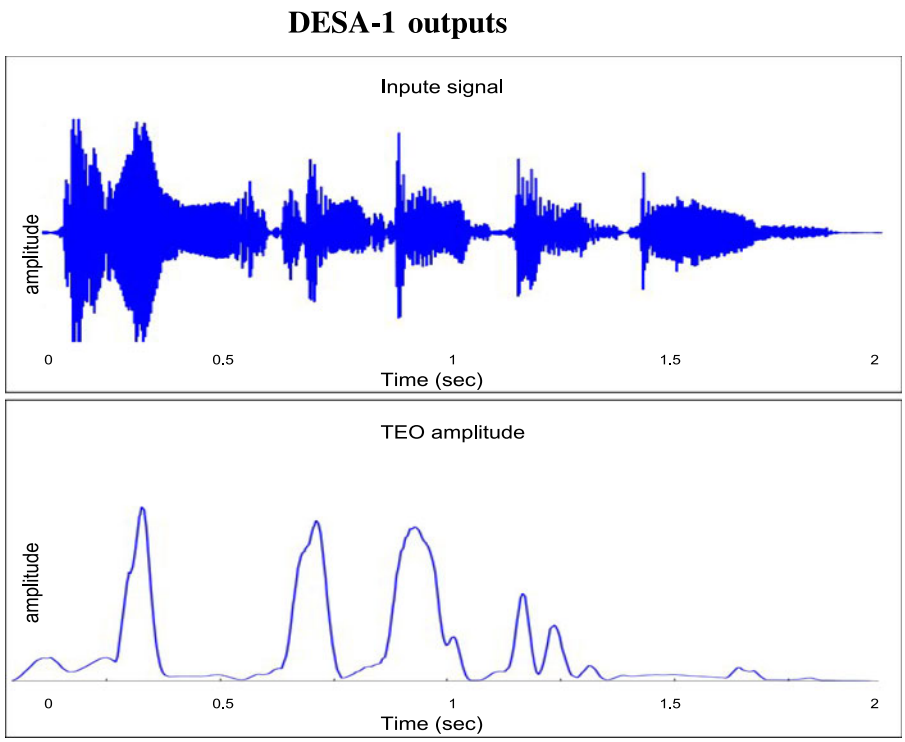

Fig. 2. Instantaneous amplitude calculated using DESA-1 from a speech sub-band. Top: Input signal. Bottom: Signal amplitude component estimated using TEO, (2).

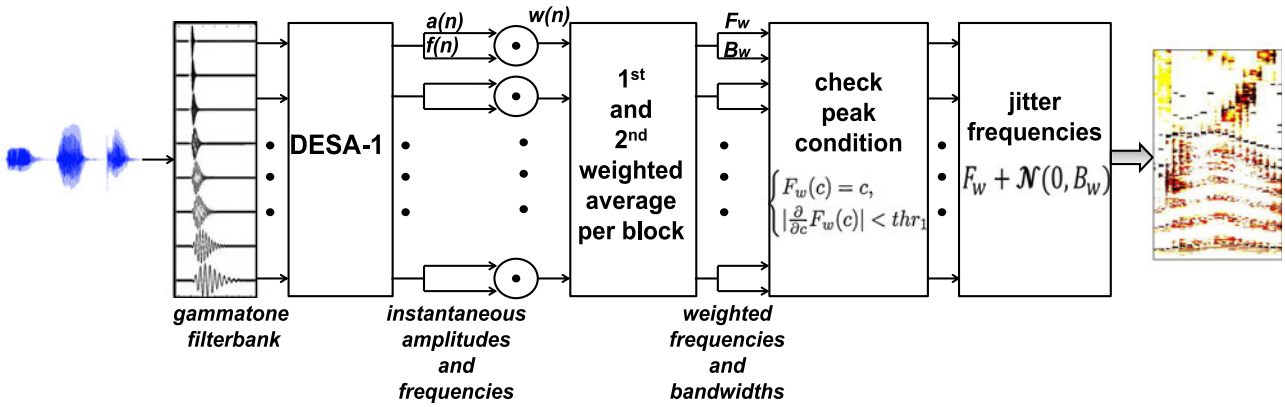

Fig. 3. Pyknogram extraction block-diagram.

amplitudes as weights. $\dot{a}(n)$ is the first difference of $a(n)$ (i.e., $a(n) - a(n-1)$). The per-frame values of F_w provide initial estimates of spectrogram peaks. This results in a time-frequency, t-f, representation of the overall signal.

In [36], the bandwidths, B_w, defined in (5) are used for analysis purposes. Here, we use them in overlap detection systems to determine the reliabilitiy of $t - f$ units. Our assumption is that large Pyknogram bandwidths correspond to higher uncertainty in frequency estimates. We investigate this in following sections by adding an uncertainty term to our frequency estimate proportional to the estimated bandwidth:

$$\tilde{F}_w(t,i) = F_w(t,i) + \epsilon_t, \tag{6}$$

where

$$\epsilon_t^i \sim \mathcal{N}(0, B_w(t,i)). \tag{7}$$

B. Pyknogram Extraction - Frequency Selection

In the second step of Pyknogram extraction, dominant harmonic peaks are selected by comparing the average frequency estimates with filter-bank center frequencies. According to [36], points at which filter-bank center frequencies coincide with the weighted frequency estimates from (4) are more reliable in estimating spectrogram peaks. In Section II-A, we showed how resonant frequencies are estimated using Teager energy operators. A considerable number of these frequencies can be omitted from the list of candidate frequencies. The assumption being that frequency estimates are more accurate when resonances align with a filter in the filter-bank. This defines the condition through which initial F_w values are tested to detect whether they correspond to prominent peaks. At frame t:

$$F_w(t,i) = F_c(i) \quad \Longleftrightarrow \quad \{i \in peaks\} \tag{8}$$

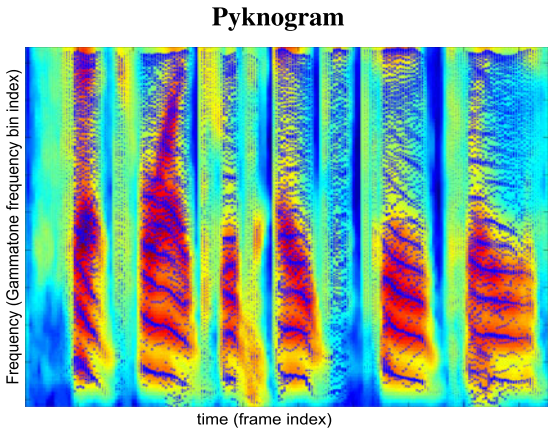

Fig. 4. Pyknogram for a given speech signal. The spectrogram is plotted in the background for comparison for a frequency range of 0-4 kHz. Pyknogram markers have been scaled by the amplitudes of corresponding t-f units. Frequencies are scaled to equivalent rectangular bandwidth (ERB) rate.

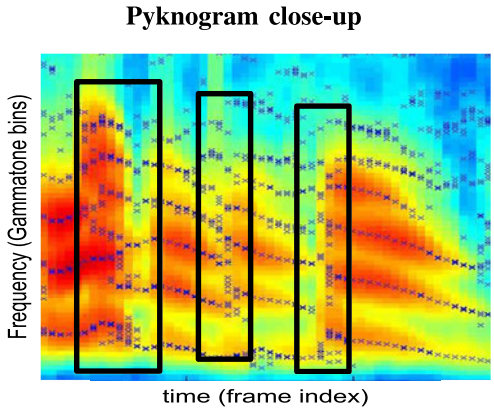

Fig. 5. A closer look on Pyknograms for overlapped speech (smaller frequency range compared to Fig. 4). The enclosed patches show discontinuities that occur in the presence of an interfering speaker.

where $F_c(i)$ is the center frequency of the ith filter in the gammatone filter-bank. Note that center frequencies are distributed in a logarithmic scale. Another peak selection condition (as shown in Fig. 3) is to limit the relative variance of selected frequencies with respect to center frequencies.

$$\left|\frac{\partial F_w(t,i)}{\partial i}\right| \approx \left|\frac{F_w(t,i+1)-F_w(t,i)}{(i+1)-i}\right| < thr, \qquad (9)$$

This condition limits non-harmonic anomalies that break the patterns in regular speech harmonics. Since such patterns are frequently observed in overlapped data, we omit this restriction from the peak-picking step.

One of the advantages of the peak-picking constraint in (8) is the quantization of spectrograms onto filter-bank center frequencies. This allows the mapping of all signals onto a unified space defined by the filter-bank, which enables reliable comparison within the time-frequency space.

Using an energy operator based approach helps avoid assumptions on the number of speakers in the signal. AM-FM decomposition is suitable since it relies on signal resonances and does not restrict signals to a specific structure or number of speakers (as opposed to models such as linear prediction). The final time-frequency representation is called a Pyknogram, $S_{pyk}(t,i)$, which is a function of time (t) and frequency index (i). $S_{pyk}(t,i)$ is obtained by applying a binary mask to the gammatone time-frequency amplitudes estimated, $A(t,i)$, from (2) in the following manner:

$$A(t,i) = \frac{1}{T}\sum_{n_t}^{n_t+T-1} a_i^2(n), \qquad (10)$$

The binary mask that results in $S_{pyk}(t,i)$ uses only amplitude values that are selected from (8) and (9).

$$S_{pyk}(t,i) = \begin{cases} A(t,i), & \text{if } F_w(t,i) \text{ satisifies (8) \& (9).} \\ 0, & \text{otherwise.} \end{cases} \qquad (11)$$

Fig. 4 shows the binary mask and underlying gammatone amplitude estimates for a given speech sample.

Using Pyknograms, we would like to investigate overlap detection methods. Discontinuities in the Pyknogram layout is an indication of interfering speech. An analogy for speech harmonic patterns are skiing tracks left behind on a snowy surface. In the single-speaker case, the patterns leave parallel tracks that progress relatively slowly over time and correspond to fundamental frequency harmonic tracks. In the presence of an interfering speaker, these patterns are distorted by similar but intersecting tracks, which adds sudden jumps along the time axis (as shown in Fig. 5 describing Pyknogram extraction). Since speakers are only capable of producing one fundamental frequency at each time instance, it is expected that the harmonic tracks should be consistent across time. This keeps harmonics parallel over short time intervals. The presence of a second speaker creates harmonic tracks that in general do not follow the same patterns, hence discontinuities are observed along time in Pyknograms. We use variations across adjacent frames as our measure of overlapped speech.

C. Unsupervised Overlap Detection

This section introduces Pyknograms as a feature extraction method for overlap detection in an unsupervised framework. As defined in Sections II-A and II-B, Pyknograms could be used to represent the signal acoustic space (similar to traditional MFCC or PLP features). Consequently, Pyknogram representation could also be a way to model the acoustic space in a hidden Markov model (HMM) topology for overlap/single-speaker/non-speech detection [38]. However, the focus here is solely on Pyknograms and their contribution to unsupervised overlap detection. In other words, we are more interested in the physical attributes of overlap vs. signle-speaker speech rather than modeling their temporal dynamics. The unsupervised approach presented in this study is to calculate the average Euclidean distance between consecutive frames across all frequencies. These distances can be used to detect sudden jumps in Pyknograms along time; much like the technique used for spectral flux estimation [39]. The distance function, D_{ovl}, at frame t is computed as the 2-$norm$ distance between consecutive

Pyknogram frames, $S_{pyk}(t,i)$ and $S_{pyk}(t-1,i)$.

$$D_{ovl}(t) = \sqrt{\sum_i \left(\left(S_{pyk}(t,i) - S_{pyk}(t-1,i) \right)^2 \right)} \quad (12)$$

which results in a frame-based value for overlap distances. Further in this study, we will investigate using longer time windows by averaging D_{ovl} of adjacent frames.

Overlapped segments are expected to have higher D_{ovl} values as compared to single-speaker speech. Fig. 5 shows instances where sudden jumps are observed in the pyknogram of an overlapped signal. The average value of these distances for all frames in a speech segment corresponds to the amount of overlapped regions (higher values are associated with greater overlap).

We evaluate the performance of our proposed detection metric on overlapped speech on the GRID database [40] (see Section III for more details on GRID). A key factor that determines the difficulty of detecting the presence of overlapped speech is the signal-to-interference (SIR) value. SIR is formally defined as the average energy of the foreground speaker to the energy of the background speaker, in dB. Of course, in the case of overlap detection, we do not favor one speaker as foreground over the other. Therefore absolute SIR values are used in this study for overlap detection. Greater absolute SIR corresponds to regions where one of the speakers has lower impact on the signal energy. Therefore it is more difficult to detect the occurrence of overlap in signals as the absolute SIR increases from zero. From here on after, when we use SIR, we imply its absolute value.

Another important factor in detecting overlap is that the SIR value will change across different frames within a single file, which is due to the non-stationary nature of speech. This poses major restrictions on the effectiveness of overlap detection evaluation, since providing frame-based ground-truth becomes unrealistically difficult. One must therefore rely on ensemble measurements over complete speech files for which the average SIR is known. We therefore introduce the segment-based D_{ovl} value which is the ensemble average of all frames within M second intervals.

$$^{M}D_{ovl}(t) = \sum_{t=t}^{t+\lfloor \frac{M}{T_s} \rfloor} D_{ovl}(t) \quad (13)$$

where T_s is the length of frame shift (in seconds), used here to determine the number of frames in an M second interval. This notion is illustrated in Fig. 6, where D_{ovl} distributions (histograms) extracted on a per-frame basis are compared with ensemble D_{ovl} distributions associated with longer durations (2 seconds). D_{ovl} for 2 second samples is calculated by averaging per-frame values. The "scores" (D_{ovl} values) in Fig. 6 are pyknogram distances calculated using (12). The top figure (Fig. 6(a)), shows the distribution of scores per *frame* (i.e. 25 msec intervals) for overlapped (target) and clean (non-target/single-speaker) data. Fig. 6(b) shows the ensemble score distributions (average score over all frames in 2 second segments). The task in overlap detection is to separate the two classes in each plot (dark blue from light blue). As observed in these distributions, the per-frame classes are almost indistinguishable (Fig. 6(a)), while in Fig. 6(b) the classes show much better separation.

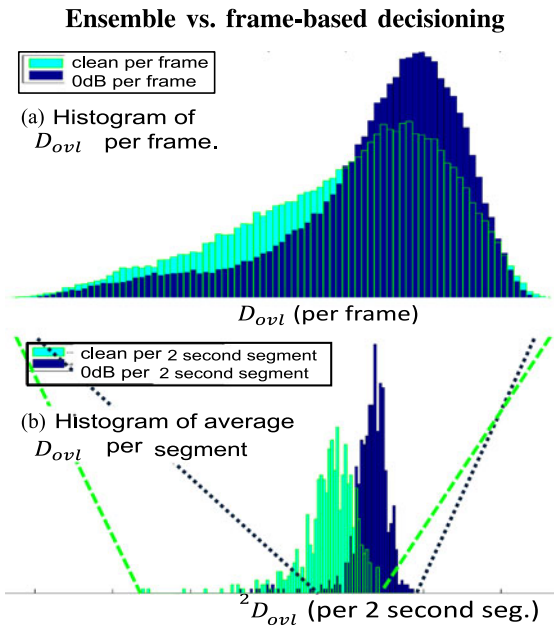

Fig. 6. The effect of ensemble decisioning on distinguishability of overlapped regions. a) shows score per frame histograms and b) shows the histogram of ensemble scores. Using multiple frames to make a decision helps separate the distributions of clean and overlapped segments.

The observation in Fig. 6 a characteristic of all non-stationary interferences, that frame-level detection is significantly less reliable compared to ensemble decisions. Therefore, longer durations should be used to detect the presence of overlap. Section III-D will investigate the relation between segment lengths (in other words, number of frames) and detection performance.

III. Experiments

This section evaluates our proposed pyknogram-based overlap detection system in terms of *accuracy, robustness,* and *precision*. Evaluation tasks for each SIR category are in the form of standard binary classification problems, where target examples are from a collection of files with fixed SIR values and non-target files are clean (single-speaker) files. We measure system performance using detection equal error-rates (EER; where false-positive and false-negative errors are equal). The expectation is that the detection algorithm should be consistent across a range of SIR values (i.e. robustness). As for precision, we are interested in determining the shortest possible signal duration before overlap detection performance significantly drops (noting the observation in Fig. 6).

Bellow, a collection of overlap detection features are presented that have previously been used to detect overlapped regions [12], [14], [21]. To the best of our knowledge, overlap detection results on this database have not been reported for any of the following features, therefore we rely on our own implementations.

A. Baseline Features

1) *Speech Kurtosis*: Kurtosis has been reported as an effective measure to detect the presence of multiple speakers in overlapped signals by several studies [18], [21], [41]. It has been shown that overlapped speech exhibits lower kurtosis compared to single-speaker speech [42]. The kurtosis of a zero-mean random variable x is defined as:

$$k_x = \frac{E\{x^4\}}{(E\{x^2\})^2} \tag{14}$$

In this case x refers to speech samples in a given frame.

2) *Spectral Flatness Measure (SFM)*: The ratio of geometric to arithmetic means of spectral magnitudes across frequency within each frame [12]. For the tth frame:

$$sfm_t = \frac{\frac{1}{N}\sum_{i=1}^{N} X(i)}{\sqrt[N]{\prod_{n=1}^{N} X(i)}} \tag{15}$$

where $X(i)$ corresponds to the magnitude spectrum at frequency ith frequency bin and N is the total number of frequency bins.

3) *Spectral Autocorrelation Peak-Valley Ratio (SAPVR)*: described briefly in Section I, this feature uses the dominance of peaks in the spectral autocorrelation in each frame as a measure to detect overlaps [14].

B. Data: Monaural Speech Separation Challenge

The data used in our controlled experiments is from the monaural speech separation and recognition challenge (aka speech separation challenge (SSC)) [4]. The objective in the SSC was to permit a large-scale comparison of techniques for the overlapped speech problem [4]. Participants were asked to identify keywords in sentences spoken by a target talker when mixed into a single channel with a background talker speaking sentences of the same structure but with different content. The data used in SSC was obtained from the larger GRID corpus [43], which is a multi-talker audio-visual sentence corpus that supports computational-behavioral studies in speech perception. In our study, we only use the audio content which consists of 1000 sentences spoken by each of 34 talkers (18 male, 16 female). The sentences are structured in the following format.

<command><color><preposition><letter><number><code>

For example, "lay white at X six now".

Seven overlapped sets are available, one clean and the rest composed of sentence pairs artificially summed at 6 signal-to-interference ratios (SIR) (+6, +3, 0, −3, −6, −9 dB). Since file durations are short (typically less than 5 seconds) and the utterances contain negligible pauses, it is reasonable to consider the average SIR values, provided for each file, a fair representation of the amount of overlap. It is also safe to assume that a given file is all speech, therefore removing the need to run speech activity detection to separate speech from silence. This assumption justifies labeling a "clean" file (no overlap) as single-speaker. Alternatively, it is also reasonable to consider an entire overlapped signal double-speaker (aka overlap) (see Fig. 7). We have down-sampled all files to 8 kHz to match telephone recordings. Note that the experiments conducted in this study do not comply with the objectives of the speech separation challenge described in [40].

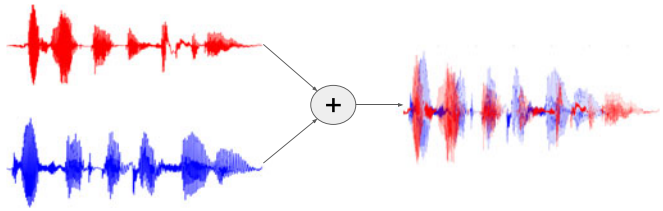

Fig. 7. Example of the mixing process for a 0 dB SIR overlapped signal. As shown on the right, it is fair to assume that overlap occurs throughout the signal.

The choice of dataset for this study is pertinent to our investigative objective, which is overlap detection. In the past, we have investigated other controlled datasets such as TIMIT [12] as well as realistic datasets including Prof-life-log [32], [33], Switchboard [6], and UT-Drive in-vehicle conversations [11]. Unfortunately, none of the aforementioned datasets are designed to contain significant amounts of overlap. One may argue that overlaps exist in conversational speech corpora such as Switchboard or the AMI meeting corpus [44]. Although many consider conversational corpora a source of overlapped speech, we argue that these corpora are not necessarily suitable for thorough and focused overlap detection analysis due to: data imbalance, varying degrees of interference (i.e., varying SIR), and non-speech vocalizations in real conversations. Data imbalance is caused since the amount of overlap is significantly less than single-speaker speech in regular conversations. Varying interference in conversational speech is due to the wide range of SIR, which makes it difficult to create ground-truth for overlap detection and thereby reduces the reliability of evaluations. Finally, a significant portion of "overlaps" in realistic speech contains laughter, which we do not intend to address in this study.

Most importantly, the amounts of overlap in "regular" (aka non-competitive conversations [10]) is not sufficient for the requirements of this study. The reason we require control over the amount of overlap in our experiments is to specifically investigate the power of our overlap detection system and not worry about target/non-target imbalances observed in conversational speech. Furthermore, the GRID corpus is isolated from variabilities other than overlapped speech, which makes it useful to study the effects of overlap. To the best of our knowledge, this dataset is the most organized publicly available corpus that contains large and controlled amounts of overlapped speech (note that we are mostly interested in *overlapped speech* and not *co-channel speech* as defined and distinguished in the introduction). Another advantage of the corpus is the fact that segments are short which makes the definition of a signal-to-interference ratio more appropriate. Had the signals been longer, say a few minutes long, the notion of a signal-to-interference ratio across the entire signal would have been less applicable, due to the non-stationary nature of speech.

TABLE I
SUMMARY OF DATA USED FOR SPEAKER RECOGNITION EXPERIMENTS

number of speakers	18 (male) 16 (female)
average file duration	1.9 (sec)
noise	interfering speakers clean, $+6$, $+3$, 0, -3, -6, -9 dB
sampling rate	8 KHz

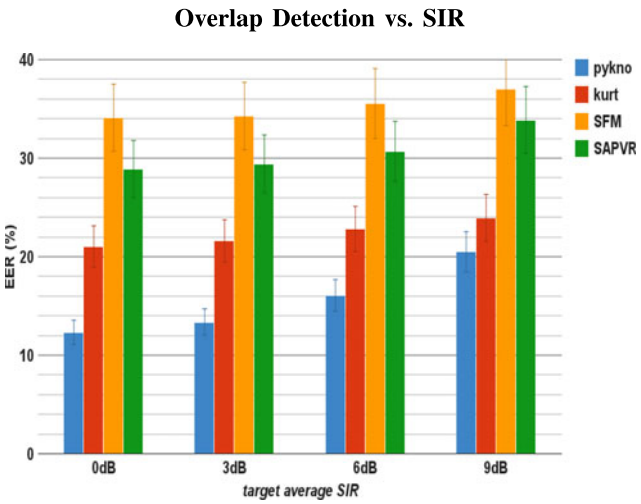

Fig. 8. Overlap detection EER for different SIR values. The higher the SIR, the more difficult it is to detect the presence of interfering speakers.

Table I shows some features of the dataset used in this study, including: number of speakers (male and female), average file duration, and overlap conditions.

C. Overlapped Speech Detection vs. SIR (Robustness & Accuracy)

Here the performance of pyknogram-based overlap detection is compared with the three baseline algorithms across different SIR values. The goal is to monitor the chances in EER as SIR values increase. The experiments are designed to compare features in terms of how much separation they can create between single-speaker from overlapped speech. Overlapped files are defined as target and single-speaker (i.e., clean) are defined as non-target files. The task is to perform binary classification using feature values as scores. The target/non-target files used in this binary classification task are obtained from a pool of overlapped and single-speaker files. This task is repeated for each SIR condition separately to monitor the impact of SIR. In each task, overlapped files with the same SIR are used as target examples and the overlap detection score (or feature value) assigned to them is compared against the scores estimated for clean files to compute the binary classification EER. Fig. 8 compares performances for the proposed and baseline systems across SIR values of $0, 3, 6$ and 9 dB.

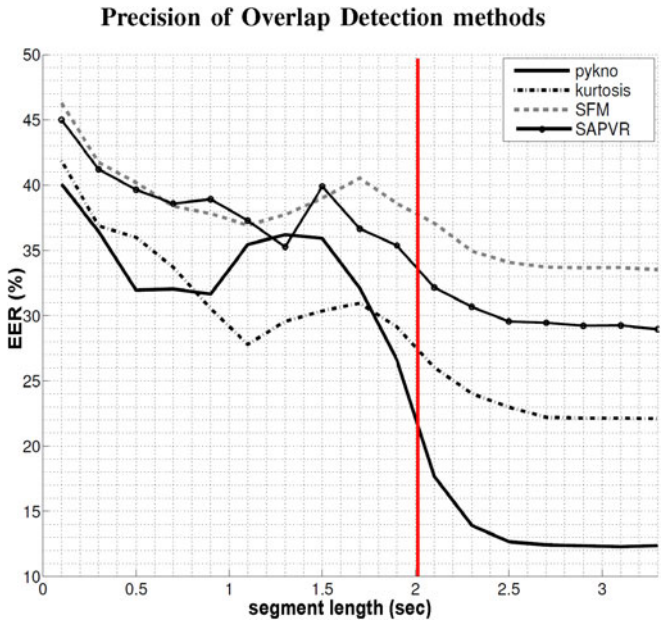

Fig. 9. Overlap detection EER as a function of segment length. The plot shows that signal lengths should be at least 2 seconds for the algorithms to start reaching their best performance.

D. Overlapped Speech Detection vs. Segment Length

A main concern in dealing with overlapped regions is that overlap decisions are less reliable as segment lengths become shorter. This restricts algorithm precision in terms of the ability to detect overlap in a frame-based framework. We recall from Section II-C, that using multiple adjacent frames in the form of average D_{ovl} (i.e., ${}^{M}D_{ovl}$) increases separability between single-speaker and overlapped speech distributions. The same averaging defined in (13) can be applied to our baseline features.

This method of treating non-stationary signals can help define the "precision" of our proposed overlap detection algorithms. In other words, what is the least number of frames required to maintain stable detection accuracy?

Precision is most valuable in tasks such as speaker diarization in conversational speech, where overlap mostly occurs at speaker transitions in turn-takings. The goal of this analysis is to evaluate system precision and compare pyknogram-based detection with baseline features. In other words, how short can overlap segments get before observing a significant drop in system performance. Once again, overlap detection performance is measured through the detection EER. Fig. 9 shows the change in system performance as shorter duration segments are used to obtain overlap decisions. It is shown that regardless of the feature used to detect overlaps, performance drops as fewer frames (shorter time segments) are used to make decisions. Furthermore, we see that performance stabilizes for all features for segments 2 seconds and longer.

IV. CASE STUDY: SPEAKER VERIFICATION IN OVERLAPPED SPEECH SIGNALS

Overlapped speech is a common phenomenon in audio recordings that are used in speaker recognition tasks. In this

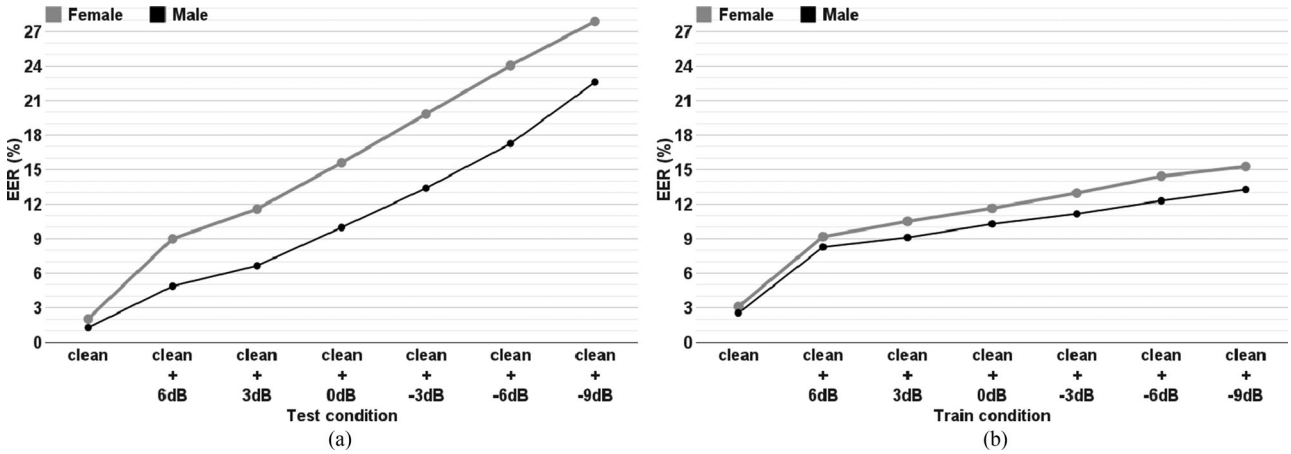

Fig. 10. The rise in EER values as we increase the effect of overlapped speech (via decreasing the SIR). Starting from clean (i.e. single-speaker speech) to lower SIR values. a) Shows the case where train files are clean, but test files contain overlaps. b) clean test files but train files contain overlaps.

section, in order to show the detrimental effects of adding overlapped data to speaker recognition, we present a case study of speaker verification on data from the monaural speech separation challenge [4]. Since most speaker recognition applications are focused on spontaneous (as opposed to text-dependent) speech, a large portion of the data are recorded from telephone or face-to-face conversations, which are prone to overlap. Examples of overlapped speech vary from instances as short as back-channeling (such as filled pauses, "aha") in a regular conversation to intentional long duration overlaps used to hold the ground in arguments, which clearly has a more substantial impact on verification accuracy. In [45], Shriberg *et al*. provide an analysis of the amount of overlapped speech in Switchboard and other corpora comprised of conversational speech. Based on the criteria used in their work (derived for automatic speech recognition purposes), 12% of words are considered overlapped in Switchboard, contributing to a large portion of the database. The frequency of overlap, however, is merely one of the factors contributing to speaker verification performance. For example, here we show that placing overlaps in train vs. test data also plays a significant role in determining system performance.

The verification experiments use 12 dimensional MFCC features (13 excluding the 0th coefficient) plus Δ and $\Delta\Delta$, which adds to a total of 36 dimensional features. 512 mixtures were used to form the Universal background model (UBM). Each speaker's Gaussian mixture model (GMM) was obtained through MAP adaptation of the means.

As mentioned above, trials are generated from train and test sets designed for the speech separation challenge. The amount of clean training data for each speaker is approximately 15 minutes. Test data are provided in six SIR conditions, which are evaluated separately (see Section IV-A). The challenge also provides overlapped training data. Overlapped training data are used in Section IV-B to train speaker models with the main speaker (i.e., model speaker in each train file) as the primary speaker and interfering speech from another randomly selected speaker. Experiments are gender-dependent, therefore the number of female speakers and male speakers is slightly different (see Table I). In total over 10000 trials are used to calculate equal error rates for each SIR condition presented in Sections IV-A and IV-B with a target to non-target ratio of 0.001.

A. Overlaps in Test Data

As a comparison benchmark, we first evaluate performance under clean train and test conditions on the SSC data. Gaussian mixture models (GMM) are adapted from a Universal back model (UBM) trained on TIMIT files [46]. For each model (training) speaker, there are 500 utterances in SSC, which are all used in the training process. Test files are available in all SIR conditions. As expected, lower SIR values correspond to higher equal error rates. The presence of a secondary speaker, clearly causes confusion in the score distribution, leading to less separability between target and imposter trials. recognition performance under clean test files and those with average SIR ranging in $+6, +3, 0, -3, -6, -9$ dB are provided in Fig. 10(a).

It is worth mentioning that the authors were tempted to compare these results with stationary noise experiments. However, contrary to our expectations, we observed that performances were better in the overlapped condition when compared to white Gaussian noise and speech-shaped noise interference, even for negative SIR values. We find this to be a misunderstanding caused by comparing stationary and non-stationary noise through the same measurement procedure, which is the SIR (or SNR). For a given target speech file, adding a certain amount of stationary noise will affect all frames, whereas in the case of non-stationary noise (here speech) only a portion of the frames receive non-uniform interference. This leads to incomparable results under presumably similar conditions which we decided to exclude from this study to avoid confusion.

B. Overlaps in Train Data

We also examine the effect of adding overlapped speech to train files (Fig. 10(b)). Fig. 11(a) and (b) compares the effects of adding overlapped speech in train and test files.

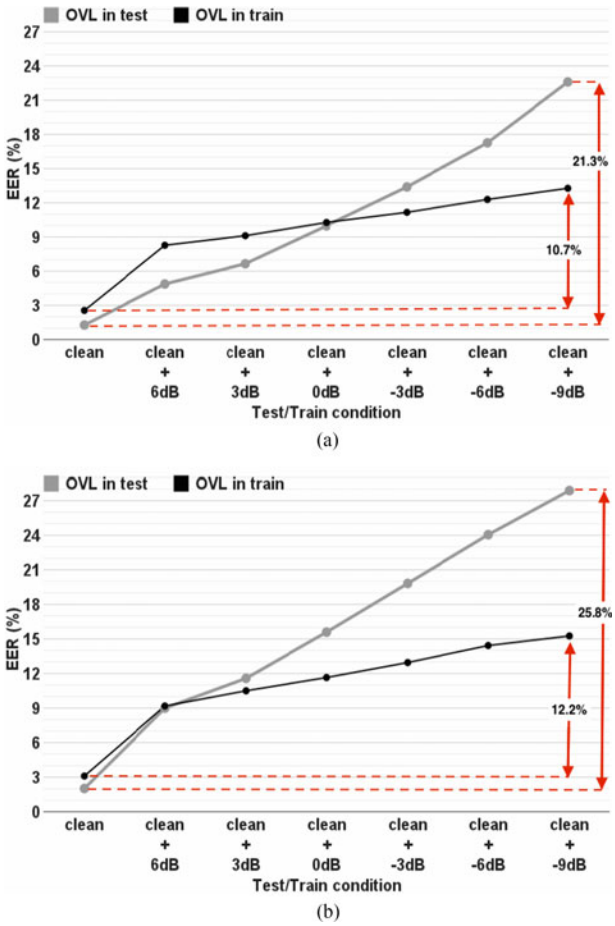

Fig. 11. Comparing the impact of increasing overlap (OVL) in train vs. test data by decreasing SIR values. Experiments for male (a) and female (b) speakers. Lower SIR drops the performance more rapidly when applied to test data.

An interesting observation is the higher rate with which the EER increases when the SIR drops for the test condition. We believe this is due to the fact that in train conditions, the training of Gaussian mixture models tends to cancel out the effect of the interfering speech. For each speaker, the GMM is trained on a set of features, some of which are influenced by the desired speaker and the rest influenced by the interfering speakers. Since multiple training files are used to model each speaker (different training files have different interfering speakers), the GMM tends to converge to a common locale in the feature space, which belongs to the speaker for whom the models are being trained. We call this effect averaging out (or cancelling out) of the interfering speakers. This to some extent slows the growth in EER as the data becomes noisier in train files. Such cancellation, however, does not exist across test files.

V. Overlap Detection Scores as Meta-Data for Speaker Verification

Using meta-data to yield more accurate decisions is a common practice in speaker verification evaluations [47], [48]. Incorporating quality measures such as speech activity detection (SAD) and effective file durations can significantly improve verification performance [48], [49] regardless of system architecture (be it i-vector, GMM-UBM, or any other system). Meta-data provides lower-level scores that help increase the distinguishability between target/imposter trials. In this study, part of the confusion in score distribution is caused by the presence of interfering speakers. We, therefore, use the scores from overlap detection algorithm(s) as secondary information to improve overall speaker verification performance.

There are several approaches through which quality-measures can be applied in a binary classification scenario [47], [50], [51]. Here, we use a stacking approach, called Q-stack, in which the quality measures (here overlap decisions) are concatenated ("stacked") with speaker verification decisions [50]. The resulting vector is a high-dimensional score vector which allows more separability due to the additional information provided by the stacked dimensions. The stacked score vectors are then processed with a support vector machine (SVM) classifier. SVM parameters are trained using a development set extracted from a separate subset of the data. In our experiments, the development set consisted of $10,000+$ trials, a quarter of which were clean trials and the remaining $7,500+$ trials contained overlapped test files with $0, 3, 6\, dB$ SIR levels. An evaluation set of size 18,000 trials with similar specifics and target-imposter ratio was used to test overall system performance.

Table II shows the improvements obtained by using the overlap detection scores individually and in combination groups. The other two features, kurtosis and SFM, show less correlation, however provide significant complementary information when combined and used alongside SAPVR and pyknogram features. The best result is obtained when all four features are concatenated, since each overlap detection system may yield better performance in certain scenarios.

The authors suggest that better individual performances from SAPVR and SFM is because of the nature of their definition which makes them superior in distinguishing harmonic structures. Since speaker identities are mostly influenced by voiced speech, this assists the speaker recognition task in quantifying the amount of voiced speech. Pyknogram-based detection is designed to locate harmonic discontinuities as opposed to the presence of harmonics.

Our experiments show that the best performance is obtained using an SVM with a radial basis function (RBF) kernel. The SVM parameter(s) (here γ) were determined through cross-validation on the development set. Class weights (i.e., target/imposter weights for the SVM classifier) and the cost (aka slack) parameter were selected according to the DCF parameters (C_{fa}, C_{miss}, and $prior$, [47]) used throughout experiments.

We also conducted an experiment using ideal overlap labels (labels from ground-truth) in the Q-stack paradigm which resulted in an upper bound in performance of 8.74% EER (23% relative improvement). We note that for the Q-stack algorithm, the relative drop in EER from using all overlap features is approximately 20%, which is not far off from when ground-truth labels are used. This confirms the effectiveness of the selected overlap detection features/scores.

TABLE II
SPEAKER VERIFICATION PERFORMANCE (EER) WITH AND WITHOUT OVERLAP DETECTION SCORES AS META-DATA

raw GMM/UBM scores	pykno	kurtosis	SFM	SAPVR	***EER (%)***
✓					11.36
✓	✓				10.19
✓		✓			13.51
✓			✓		28.35
✓				✓	9.48
✓	✓	✓			10.20
✓	✓		✓		10.47
✓	✓			✓	9.57
✓	✓	✓	✓		10.31
✓	✓	✓		✓	9.18
✓	✓	✓	✓	✓	**9.10**

Grey cells highlight the features used in each experiment. The relative change in EER is presented in the last column.

VI. CONCLUSION

This study provided an analysis of overlapped speech detection by proposing a novel approach to identify overlapped from single-speaker speech. The proposed method was based on harmonically enhanced spectrograms, called Pyknograms, which are the result of applying a binary mask of the harmonic structure onto the time-frequency power spectrum. The binary mask uses Teager-Kaiser energy operators to estimate resonant frequencies in the signal, which are then mapped onto a time-frequency units. Pyknograms were found useful in distinguishing overlapped from single-speaker speech, due to differences in their dynamic structures. A Euclidean distance measure of adjacent time units was used to quantify Pyknogram dynamics and used as an overlap detection feature. Pyknogram-based overlap detection was compared with 3 existing baseline feature in an unsupervised framework. We compared the features in terms of robustness by evaluating system performance under different signal-to-interference ratios (SIR). Additionally, features were compared in terms of precision by calculating detection error rates under different signal durations. One of the contributions of this study is to point out that as a non-stationary form of interference, overlap detection is more viable when long signal durations are used. We test this by using multiple frames to compute features for the overlap detection system. Results show promising results for Pyknogram-based unsupervised overlap detection. Future work in this regard should include supervised temporal modeling of Pyknograms for overlap detection in co-channel conversational or meeting speech. A supervised algorithm may act as a speaker count measure for a diarization system.

This study also investigated the impact of overlapped speech on various aspects of speaker recognition. The analysis includes a comparison of the effect of introducing overlap to test vs. training data in a speaker verification framework. Results show a trend that interferer variability in training data contributes to an averaging affect that lowers the impact of overlap speech on speaker verification performance. The final section considered using overlapped detection results as meta-data for a given speaker verification task. The meta-data was incorporated using the Q-stack algorithm and a support vector machine (SVM) classifier to improve verification performance by taking advantage of a high-dimensional score-space. We established a lower bound of the achievable equal-error-rate for the Q-stack algorithm by calculating the results using ground-truth overlapped labels, which yields a 23% relative improvement. Using the proposed overlap detection system and other existing features the relative improvement was 20%, a mere 3% lower than the best achievable performance provided by the lower bound.

REFERENCES

[1] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 407–424, Sep. 1997.

[2] T. Quatieri and R. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 1, pp. 56–69, Jan. 1990.

[3] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multipitch estimation using the em algorithm for co-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1993, pp. 728–731.

[4] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 1–15, 2010.

[5] R. E. Yantorno, "Co-channel speech and speaker identification study," Final report for Summer Research Faculty Program, Air Force Office of Scientific Research, Speech Processing Lab, Rome Labs, New York, 1998.

[6] N. Shokouhi and J. H. Hansen, "Probabilistic linear discriminant analysis for robust speaker identification in co-channel speech," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3016–3020.

[7] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved diarization in multiparty meetings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, USA, 2008, pp. 4353–4356.

[8] M. Zelenak, C. Segura, J. Luque, and J. Hernando, "Simultaneous speech detection with spatial features for speaker diarization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 436–446, Feb. 2012.

[9] S. H. Yella and H. Bourlard, "Improved overlap speech diarization of meeting recordings using long-term conversational features," in *Proc. 2013 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7746–7750.

[10] E. A. Schegloff, "Overlapping talk and the organization of turn-taking for conversation," *Lang. Soc.*, vol. 29, no. 1, pp. 1–63, 2000.
[11] A. Sathyanarayana, N. Shokouhi, S. O. Sadjadi, and J. H. Hansen, "Belt up: Investigating the impact of in-vehicular conversation on driving performance," in *Proc. IEEE Int. Veh. Symp. (IV)*, 2013, pp. 1071–1076.
[12] N. Shokouhi, A. Sathyanarayana, S. Sadjadi, and J. H. L. Hansen, "Overlapped-speech detection with applications to driver assessment for in-vehicle active safety systems," in *Proc. IEEE Inter. Conf. Acoust., Speech, Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 2834–2838.
[13] B. Smolenski and R. Ramachandran, "Usable speech processing: A filterless approach in the presence of interference," *IEEE Circuits Syst. Mag.*, vol. 11, no. 2, pp. 8–22, Apr.–Jun. 2011.
[14] K. Krishnamachari, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions," in *Proc. IEEE Int. Symp. Intell. Signal Process. Commun. Syst.*, Nov. 2000, pp. 710–713.
[15] J. M. Lovekin, K. R. Krishnamachari, and R. E. Yantorno, "Adjacent pitch period comparison as a usability measure of speech segments under co-channel conditions," in *Proc. IEEE Int. Symp. Intell. Signal Process. Commun. Syst.*, Apr. 2001, pp. 139–142.
[16] M. Wu, D. L. Wang, and G. J. Brown, "A multi-pitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 229–241, May 2003.
[17] Y. Shao and D. L. Wang, "Co-channel speaker identification using usable speech extraction based on multi-pitch tracking," in *Proc. IEEE Inter. Conf. Acoust., Speech, Signal Process.*, Hong Kong, 2003, pp. 205–208.
[18] S. N. Wrigley, G. J. Brown, W. Vincent, and S. Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Trans. Audio Speech Lang. Process.*, vol. 13, no. 1, pp. 84–91, Jan. 2005.
[19] P. Dighe, M. Ferras, and H. Bourlard, "Detecting and labeling speakers on overlapping speech using vector taylor series," in *Proc. INTERSPEECH*, 2014, pp. 592–596.
[20] N. Shokouhi, S. O. Sadjadi, and J. H. Hansen, "Co-channel speech detection via spectral analysis of frequency modulated sub-bands." in *Proc. INTERSPEECH*, 2014, pp. 2380–2384.
[21] K. Boakye, "Audio segmentation for meeting speech processing," Ph.D. dissertation Univ. Calif., Berkeley, CA, USA, 2008.
[22] K. Boakye, O. Vinyals, and G. Friedland, "Improved overlapped speech handling for speaker diarization," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 4353–4356.
[23] J. T. Geiger, R. Vipperla, S. Bozonnet, N. Evans, B. Schuller, and G. Rigoll, "Convolutive non-negative sparse coding and new features for speech overlap handling in speaker diarization," *Energy*, vol. 40, pp. 1–5, 2012.
[24] M. Giuliani, T. L. Nwe, and H. Li, "Meeting segmentation using two-layer cascaded subband filters," in *Proc. 5th Int. Conf. Chin. Spoken Lang. Process.*, 2006, pp. 672–682.
[25] P. Maragos, J. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3024–3051, Oct. 1993.
[26] J. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Proc. IEEE Inter. Conf. Acoust., Speech, Signal Process.*, Apr. 1990, vol. 1, pp. 381–384.
[27] P. Maragos, A. Potamianos, R. Potamianos, B. Santhanam, and G. Xx, "Instantaneous energy operators: Applications to speech processing and communications," in *Proc. IEEE Workshop Nonlinear Signal Image Proc.*, Thessaloniki, Greece, 1995, pp. 955–958.
[28] Y. Litvin, I. Cohen, and D. Chazan, "Monaural speech/music source separation using discrete energy separation algorithm," *Signal Process.*, vol. 90, no. 12, pp. 3147–3163, 2010.
[29] W. Lin, C. Hamilton, and P. Chitrapu, "A generalization to the Teager-Kaiser energy function and application to resolving two closely-spaced tones," in *Proc. IEEE Inter. Conf. Acous., Speech, Signal Process.*, May 1995, vol. 3, pp. 1637–1640.
[30] X. Hu, S. Peng, and W.-L. Hwang, "Multicomponent AM-FM signal separation and demodulation with null space pursuit," *Signal, Image Video Process.*, vol. 7, pp. 1093–1102, 2012.
[31] B. Santhanam and P. Maragos, "Multicomponent AM-FM demodulation via periodicity-based algebraic separation and energy-based demodulation," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 473–490, Mar. 2000.
[32] N. Shokouhi, A. Ziaei, A. Sangwan, and J. H. L. Hansen, "Robust overlapped speech detection and its application in word-count estimation for prof-life-log data," in *Proc. IEEE Inter. Conf. Acoust., Speech, Signal Process.*, Brisbane, Australia, Apr. 2015, pp. 4724–4728.
[33] A. Ziaei, A. Sangwan, L. Kaushik, and J. H. Hansen, "Prof-life-log: analysis and classification of activities in daily audio streams," in *Proc. 2015 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4719–4723.
[34] K. Boakye, O. Vinyals, and G. Friedland, "Two's a crowd: Improving speaker diarication by automatically identifying and excluding overlapped speech," in *Proc. INTERSPEECH*, Brisbane, Australia, 2008, pp. 32–35.
[35] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1995, vol. 1, pp. 784–787.
[36] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3795–3806, Jun. 1996.
[37] L. Cohen and C. Lee, "Instantaneous bandwidth for signals and spectrogram," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1990, vol. 5, pp. 2451–2454.
[38] G. Saon, S. Thomas, H. Soltau, S. Ganapathy, and B. Kingsbury, "The IBM speech activity detection system for the darpa rats program." in *Proc. INTERSPEECH*, 2013, pp. 3497–3501.
[39] S. Rossignol and O. Pietquin, "Single-speaker/multi-speaker co-channel speech classification," in *Proc. INTERSPEECH*, 2010, pp. 2322–2325.
[40] M. Cooke and T. Lee. Speech separation challenge, 2006. [Online]. Available: http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge
[41] K. Krishnamachari, R. E. Yantorno, J. M. Lovekin, D. S. Benincasa, and S. J. Wenndt, "Use of local kurtosis measure for spotting usable speech segments in co-channel speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Salt Lake City, UT, USA, 2001, pp. 649–652.
[42] J. LeBlanc and P. De Leon, "Speech separation by kurtosis maximization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1998, vol. 2, pp. 1029–1032.
[43] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 120, no. 5, pp. 2421–2424, Nov. 2006.
[44] J. Carletta *et al.*, "The AMI meeting corpus: A pre-announcement," in *Proc. 2nd Int. Conf. Mach. Learn. Multimodal Interaction* (ser. MLMI'05), Berlin, Heidelberg: Springer-Verlag, 2006, pp. 28–39.
[45] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Proc. Eurospeech*, 2001, pp. 1359–1362.
[46] S. Sadjadi, M. Slaney, and L. Heck, "MSR identity toolbox v1.0: A matlab toolbox for speaker-recognition research," *Speech Lang. Process. Tech. Committee Newsletter*, Nov. 2013. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=205119
[47] N. Brummer and E. De Villiers. BOSARIS toolkit, 2011. [Online]. Available: http://sites.google.com/site/bosaristoolkit
[48] M. Mandasari, R. Saeidi, M. McLaren, and D. van Leeuwen, "Quality measure functions for calibration of speaker recognition systems in various duration conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 11, pp. 2425–2438, Nov. 2013.
[49] T. Hasan, S. Sadjadi, G. Liu, N. Shokouhi, H. Boril, and J. Hansen, "CRSS systems for 2012 NIST speaker recognition evaluation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 6783–6787.
[50] K. Kryszczuk and A. Drygajlo, "Improving biometric verification with class-independent quality information," *IET Signal Process.*, vol. 3, no. 4, pp. 310–321, Jul. 2009.
[51] F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification in score-ageing-quality classification space," *Comput. Speech Lang.*, vol. 27, no. 5, pp. 1068–1084, 2013.

Navid Shokouhi received the Bachelor's degree in electrical engineering in 2011 from Amirkabir University of Technology (Tehran Polytechnique), Tehran, Iran. He received the Ph.D. degree in electrical engineering from the University of Texas, Dallas, USA, in 2016. His main research involves applying Machine Learning and Signal Processing Techniques in Spoken Language Technology. His Ph.D. research focused on multispeaker speaker recognition and Diarization, which requires developing robust speaker recognition algorithms for co-channel signals.

John H. L. Hansen (S'81–M'82–SM'93–F'07) received the Ph.D. and M.S. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, Georgia, in 1988 and 1983, and B.S.E.E. degree from Rutgers University, College of Engineering, New Brunswick, NJ, USA, in 1982. He received the honorary degree Doctor Technices Honoris Causa from Aalborg University, Aalborg, Denmark, in April 2016, in recognition of his contributions to speech signal processing and speech/language/hearing science. He joined University of Texas at Dallas (UTDallas), Erik Jonsson School of Engineering and Computer Science in 2005, where he is currently serving as Jonsson School Associate Dean for Research, as well as a Professor of electrical engineering, the Distinguished University Chair in Telecommunications Engineering, and a joint appointment as Professor in the School of Behavioral and Brain Sciences (Speech and Hearing). He previously served as Department Head of Electrical Engineering from August 2005 to December 2012, overseeing a +4x increase in research expenditures ($4.5 to 22.3 M) with a 20% increase in enrollment along with hiring 18 additional T/TT faculty, growing UTDallas to the 8th largest EE program from ASEE rankings in terms of degrees awarded. At UTDallas, he established the Center for Robust Speech Systems (CRSS). Previously, he served as Department Chairman and Professor of Department of Speech, Language, and Hearing Sciences (SLHS), and Professor in Department of Electrical and Computer Engineering, at University of Colorado Boulder (1998–2005), where he co-founded and served as an Associate Director of the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTDallas. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human-computer interaction. He has been named IEEE Fellow (2007) for contributions in "Robust Speech Recognition in Stress and Noise," International Speech Communication Association (ISCA) Fellow (2010) for contributions on research for speech processing of signals under adverse conditions, and received The Acoustical Society of Americas 25 Year Award (2010) in recognition of his service, contributions, and membership to the Acoustical Society of America. He is currently serving as the elected Vice-President of ISCA and a member of the ISCA Board. He was also selected and is serving as Vice-Chair on U.S. Office of Scientific Advisory Committees (OSAC) for OSAC-Speaker in the voice forensics domain (2015–2017). Previously he served as IEEE Technical Committee (TC) Chair and Member of the IEEE Signal Processing Society: Speech-Language Processing Technical Committee (SLTC) (2005–2008; 2010–2014; elected IEEE SLTC Chairman for 2011–2013, Past-Chair for 2014), and elected ISCA Distinguished Lecturer (2011/2012). He has served as a Member of IEEE Signal Processing Society Educational Technical Committee (2005–2008; 2008–2010); Technical Advisor to the U.S. Delegate for NATO (IST/TG-01); IEEE Signal Processing Society Distinguished Lecturer (2005/06), Associate Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1999), Associate Editor for IEEE SIGNAL PROCESSING LETTERS (1998–2000), Editorial Board Member for IEEE SIGNAL PROCESSING MAGAZINE (2001–2003); and Guest Editor (October 1994) for special issue on Robust Speech Recognition for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He has served on Speech Communications Technical Committee for Acoustical Society of America (2000–2003), and previously on ISCA Advisory Council. He has supervised 77 Ph.D./M.S. thesis candidates (40 Ph.D., 37 M.S./M.A.), received the 2005 University of Colorado Teacher Recognition Award as voted on by the student body, author/co-author of 640 journal and conference papers including 12 textbooks in the field of speech processing and language technology, coauthor of textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), co-editor of DSP for *In-Vehicle and Mobile Systems* (Springer, 2004), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), *In-Vehicle Corpus and Signal Processing for Driver Behavior* (Springer, 2008), and *lead author of the report The Impact of Speech Under Stress on Military Speech Technology*, (NATO RTO-TR-10, 2000). He also organized and served as General Chair for ISCA Interspeech-2002, September 16–20, 2002, Co-Organizer and Technical Program Chair for IEEE ICASSP-2010, Dallas, TX, USA, March 15–19, 2010, and Co-Chair and Organizer for IEEE SLT-2014, December 7–10, 2014 in Lake Tahoe, NV, USA.