Prosodic and other Long-Term Features for Speaker Diarization

Gerald Friedland, *Member, IEEE*, Oriol Vinyals, Yan Huang, *Member, IEEE*, and Christian Müller

***Abstract*—Speaker diarization is defined as the task of determining "who spoke when" given an audio track and no other prior knowledge of any kind. The following article shows how a state-of-the-art speaker diarization system can be improved by combining traditional short-term features (MFCCs) with prosodic and other long-term features. First, we present a framework to study the speaker discriminability of 70 different long-term features. Then, we show how the top-ranked long-term features can be combined with short-term features to increase the accuracy of speaker diarization. The results were measured on standardized datasets (NIST RT) and show a consistent improvement of about 30% relative in diarization error rate compared to the best system presented at the NIST evaluation in 2007.**

***Index Terms*—Long-term features, prosody, speaker diarization.**

I. Introduction

COMPARED to other disciplines where machine learning is applied, speech research has been relatively conservative regarding the use of different features. A small set of standard features, such as mel frequency cepstral coefficients (MFCCs) or perceptual linear prediction (PLP), in different dimensionalities tends to be used for almost any speech-related task even when problems seem to be orthogonal, such as speech and speaker recognition. The field of speaker diarization is no exception. Here, the task is to segment audio into speaker-homogeneous regions with the goal of answering the question, "Who spoke when?". Current systems usually rely on the combination of Gaussian mixture models (GMMs) of frame-based cepstral features [1].

In the related field of speaker recognition, task-specific features have been successfully applied in combination with MFCCs. These features are often obtained on portions of speech longer than one frame and are therefore referred to as long-term features. In this paper, we present a systematic investigation of the speaker discriminability of 70 long-term features. We provide additional evidence that despite the dominance of short-term cepstral features in speaker recognition, a number of long-term features can provide significant information for speaker discrimination. As already suggested by [2], looking at patterns derived from a larger segment of speech can reveal individual characteristics of the speakers' voices as well as their speaking behavior, which cannot be captured by exclusively using frame-based short-term cepstral analysis.

Using a combination of the top-ten ranked prosodic and long-term features combined with regular MFCCs leads to a 30% relative improvement in terms of the diarization error rate (DER). The results were measured on the U.S. National Institute of Standards and Technology Rich Transcription (NIST RT) test and evaluation datasets[1], and were compared to the top-performing system of the NIST RT evaluation in 2007.

The paper is structured as follows. Section II surveys related work in speaker diarization and the use of alternative features, i.e., non short-term noncepstral features in speaker recognition. Section III presents our baseline, namely the ICSI speaker diarization system. Section IV presents the pool of features and explains our selection framework with the applied ranking methodology. Section V discusses the actual integration of the features into the ICSI speaker diarization system and presents the experimental results on different NIST benchmarks. Section VI summarizes the article and presents future work.

II. Related Work

A. Long-Term Features and Speaker Recognition

Short-term cepstral features are generally referred to as *low-level* features reflecting the voice parameters of the speaker as opposed to *higher-level* features that capture phonetic, prosodic, and lexical information. Unfortunately, some prosodic features are very hard to compute, while others are inherently difficult to infer solely from acoustics (e.g., lip-roundness). Therefore, higher-level features have increasingly attracted attention only in the last decade.

In [2], the author summarizes approaches using higher-level information for speaker recognition, a field that is closely related to diarization, and describes them in terms of their type, temporal span, and relevance to the task. It was shown that systems using a combination of cepstral and higher-level features outperformed standard systems, especially when the amount of available training data was increased. This confirms the assumption that short-term cepstral systems generally perform well be-

Manuscript received May 23, 2008; revised December 10, 2008. Current version published June 10, 2009. This work was supported in part by IARPA VACE. The work of G. Friedland and C. Müller was supported by fellowships within the postdoctoral program of the German Academic Exchange Service (DAAD) and the work of O. Vinyals was supported by the Fundación Caja Madrid. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Richard C. Rose.

G. Friedland and O. Vinyals are with the International Computer Science Institute, Berkeley, CA 94704-1198 USA.

Y. Huang was with the International Computer Science Institute, Berkeley, CA 94704-1198 USA. She is now with Li Creative Technologies, Inc., Florham Park, NJ 07932 USA.

C. Müller was with the International Computer Science Institute, Berkeley, CA 94704-1198 USA. He is now with DFKI, D-66123 Saarbrücken, Germany.

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/TASL.2009.2015089

[1]http://www.nist.gov/speech/tests/rt

cause they reflect information about the speaker's physiology and do not rely on the phonetic content (which makes them inherently text-independent). However, long-range information that also resides in the signal is only exploited in the combined systems. In addition, as pointed out by [2], higher-level features also have the potential of increased robustness to channel variation, since lexical usage or temporal patterns do not change with the change of acoustic conditions.

Clearly, lexical idiosyncrasies are not investigated here at all. Therefore, rather than using the broader term *higher-level*, we refer to non-cepstral features as *prosodic* and *long-term* features. Following the definition proposed by [2], long-term information refers to features that are extracted over regions longer than a frame. Prosodic features capture variations in intonation, timing, and loudness that are specific to the speaker. Because such features are suprasegmental, i.e., extend beyond one segment, they can be considered a subset of *long-term* features. Here, mainly pitch and energy dynamics are investigated. However, [2] itemizes further types of prosodic features such as (explicit) syllable-based prosody sequences, interpause/conversation level statistics, and durational features. The relevant subset of studies collected by [2] has been reviewed for this paper. It is summarized in the following along with newer references.

A method for modeling the pitch contour for speaker recognition is described in [3]. The movements of the speaker's pitch are obtained by fitting a piecewise linear model to the actual pitch track. The parameters by which the stylized model is described are then used as features for speaker verification. In [4], variants of this basic approach are described in which rises and falls of the fitted pitch and energy values are extracted and modeled as bi-grams. In [5], duration, pitch, and energy statistics are modeled for units between pauses. A very long span of speech is looked at by [6], [7], where statistics are computed over an entire conversation side. Then, the distances of each conversation-level feature vector are compared for target and impostor speakers using log-likelihood ratios.

Smaller time units, namely syllables, are used in [8], [9], which have the advantage of resulting in a larger number of samples. Syllables are automatically inferred from speech recognizer output. A large number of pitch, duration, and energy values are extracted from each syllable. A set of GMMs is created for each sequence of syllables and pauses. For a given sample, the posterior probabilities of the Gaussian components are concatenated and utilized in a support vector machine.

In summary, long-term features have been investigated for several years and indications have been provided that they can be useful for speaker recognition. However, as will be shown in the next section, their use in speaker diarization has never been explored.

B. Speaker Diarization

As already explained in Section I, the goal of speaker diarization is to segment audio into speaker-homogeneous regions with the ultimate goal of answering the question, "Who spoke when?" [1]. While in speaker recognition, models are trained for a specific set of target speakers which are applied to an unknown test speaker for acceptance (target and test speaker match) or rejection (mismatch), in speaker diarization no prior knowledge about the identity or number of the speakers in the recording is given.

Conceptually, a speaker diarization system therefore performs three tasks: First, discriminate between speech and nonspeech regions (speech activity detection); second, detect speaker changes to segment the audio data; third, group the segmented regions together into speaker-homogeneous clusters. Different approaches have been developed over the years. They can be roughly organized into one-stage versus two-stage algorithms and model-based versus nonmodel-based systems.

In two-stage speaker diarization approaches, the speaker segmentation step initially detects speaker change points, and is essentially a two-way decision problem. At each point, a decision on whether this is a speaker change point or not needs to be made. Subsequently, the speech segments, each of which contains only one speaker, are then clustered using either top-down or bottom-up approaches. Many state-of-the-art speaker diarization systems, including the ICSI Speaker Diarization engine, instead use a one-stage approach—they unify the segmentation and clustering steps into a single step. A very popular method is the combination of agglomerative clustering with Bayesian Information Criterion (BIC) [10] and Gaussian mixture models of frame-based cepstral features (MFCCs) [1] (see Section III). Recently, a new speaker clustering approach, which applies the Ng–Jorden–Weiss (NJW) spectral clustering algorithm to speaker diarization has also been reported [11].

In model-based approaches, pretrained speech and silence models are used for segmentation. The decision about speaker change is made based on frame assignment, i.e., the detected silence gaps are considered to be the speaker change points. In nonmodel based approaches, no pretrained speech is used. Rather, a metric between two contiguous speech segments is defined and the decision is made via a thresholding procedure. During the last years, research has concentrated on finding metrics for speaker change detection. Examples are BIC [10], cross BIC (XBIC) [12], [13], generalized likelihood ratio (GLR) [14], Gish distance [15], Kullback–Leibler distance (KL) [16], and the divergence shape distance (DSD) [17].

As this section showed, many different machine-learning strategies have been explored in order to solve the speaker diarization problem. However, with very few exceptions, exploration of features for use in speaker diarization has been limited mostly to varying the dimensionality of the cepstral features. In [18] the authors propose a framework for combining MFCC features with PLP. The robustness improvement is minimal. In [19], the authors introduce the use of delay features, which is the delay between signals from different microphones in an array. However, these features can only be extracted under a condition where multiple distant microphones are available.

We think the exploration of different machine learning techniques to improve the performance of speaker diarization is very important and should continue to be actively pursued by the community. In addition, we believe and provide evidence in this article that investigating different front-end features is equally important because they can improve the robustness of speaker diarization, as well as provide further insights into the underlying problem of speaker discrimination.

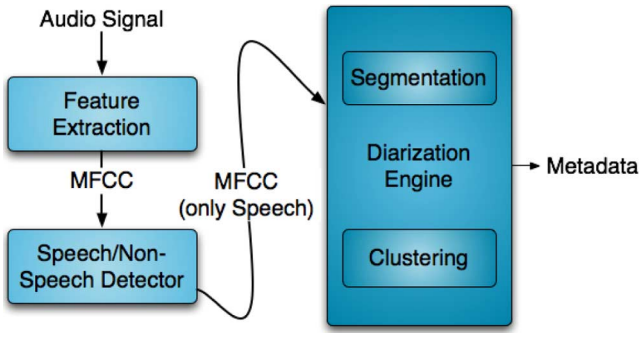


Fig. 1. Diagram illustrating the baseline ICSI Speaker Diarization Engine which is described in Section III. The audio signal, given as cepstral features (MFCC) undergoes a two stage process: Speech/Nonspeech filtering, and one-step segmentation and clustering.

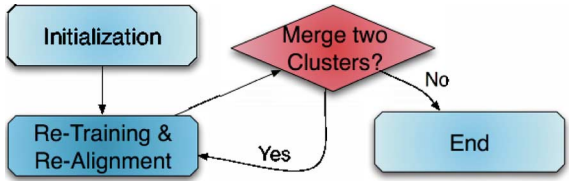


Fig. 2. Agglomerative clustering approach of the ICSI Speaker Diarization Engine as explained in Section III. Re-training and re-segmentation ends when no more models can be merged as determined by the BIC score. At the end, the number of clusters should hopefully equal the number of speakers.

III. Baseline ICSI Speaker Diarization Engine

The ICSI Speaker Diarization engine is illustrated in Fig. 1. At a high-level, the engine extracts MFCC features from a given audio track, discriminates between speech and nonspeech regions (speech activity detection), and uses an agglomerative clustering approach to perform both segmentation of the audio track into speaker-homogeneous time segments and the grouping of these segments into speaker-homogeneous clusters in one step.

The audio track is usually processed as 19th-order MFCC features using a frame size of 30 ms, with a step size of 10 ms. Speech activity regions are determined using a state-of-the-art speech/nonspeech detector [20]. The detector performs iterative training and re-segmentation of the audio into three classes: speech, silence, and audible nonspeech. To bootstrap the process, an initial segmentation is created with an hidden Markov model (HMM) trained on broadcast news data. The nonspeech regions are then excluded from the agglomerative clustering, which is explained in the following paragraph.

The algorithm is initialized using k clusters, where k is larger than the number of speakers that are assumed to appear in the recording.[2]

In order to train initial GMMs for the k speaker clusters an initial segmentation is generated by uniformly partitioning the audio into k segments of the same length. The algorithm then performs the following loop.

- Re-Segmentation: Run Viterbi Alignment to find the optimal path of frames and models. As classifications based on 30-ms frames are very noisy, a minimum duration of 2.5 s is assumed for each speech segment.
- Re-Training: Given the new segmentation of the audio track, compute new Gaussian mixture models for each of the clusters.
- Cluster Merging: Given the new GMMs, try to find the two clusters that most likely represent the same speaker. This is done by computing the BIC score (Bayesian information criterion) of each of the clusters and the BIC score of a new GMM trained on the merged segments for two clusters. If the BIC score of the merged GMM is smaller than or equal to the sum of the individual BIC scores, the two models are merged and the algorithm loops at the re-segmentation using the merged GMM. If no pair is found, the algorithm stops.

Fig. 2 illustrates the steps of the algorithm. A more detailed description can be found in [21] and [13]. As a result of different optimization approaches [22], our current implementation runs at about $0.6\times$ realtime. This means for 10 min of audio data, the engine needs roughly 6 min for diarization. Although the actual execution speed is ultimately dependent on CPU, memory, number of speakers in the meeting, speech/nonspeech ratio, number of speaker turns, and other factors.

The output consists of meta-data describing speech segments in terms of starting time, ending time, and speaker cluster name. This output is usually evaluated against manually annotated ground truth segments. A dynamic programming procedure is used to find the optimal one-to-one mapping between the hypothesis and the ground truth segments so that the total overlap between the reference speaker and the corresponding mapped hypothesized speaker cluster is maximized. The difference is expressed as Diarization Error Rate, which is defined by NIST.[3] The Diarization Error Rate (DER) can be decomposed into three components: misses (speaker in reference, but not in hypothesis), false alarms (speaker in hypothesis, but not in reference), and speaker errors (mapped reference is not the same as hypothesized speaker).

The ICSI Speaker Diarization System has competed in the NIST evaluations of the past several years and established itself well among state-of-the-art systems.[4]

The current official score is 21.74% DER for the single-microphone case (RT07 evaluation set). This error can be decomposed in 6.8% speech/nonspeech error and 14.9% speaker clustering error. The total speaker error includes all wrongly classified segments, including overlapped speech.

The following sections present our approach on improving these scores using prosodic and long-term features.

IV. Feature Selection

Table I shows the initial list of candidate prosodic and long-term features. They are extracted using a library based on

[2]Sometimes the number of speakers is known in advance because of a specific task, sometimes heuristics are used to come up with a guess. During NIST evaluations, we found that for a 30-min broadcast news snippet $k = 64$ and for meetings $k = 16$ are good choices.

[3]http://nist.gov/speech/tests/rt/rt2004/fall

[4]NIST rules prohibit publication of results other than our own. Please refer to the NIST website for further information: http://www.nist.gov/speech/tests/rt/rt2007

TABLE I
FEATURE RANKING RESULT BASED ON THE RATIO OF THE BETWEEN-SPEAKER AND WITHIN-SPEAKER VARIABILITY OBTAINED ON A PER-UTTERANCE BASIS ON THE TIMIT DATABASE

rank	feature category	feature	$\frac{between}{within}$
1	pitch	f0_median	26.375387
2	pitch	f0_mean	9.412929
3	formants	f4_stdev	3.356307
4	pitch	f0_min	3.072750
5	formants	f4_min	2.885247
6	harmonics-to-noise ratio	h2n_mean	2.845387
7	formants	f4_mean	2.615583
8	formants	f5_mean	2.437822
9	long-term average spectrum	ltas_stdev	2.322184
10	formants	f5_stdev	2.251312
11	formants	f4_median	2.191332
12	formants	f5_min	2.170989
13	formants	min_disp	2.058759
14	formants	f5_median	2.043392
15	energy	en_max	1.952964
16	formants	f3_stdev	1.517770
17	formants	mean_disp	1.516091
18	formants	f3_min	1.356388
19	formants	f3_mean	1.236645
20	long-term average spectrum	ltas_slope	0.987187
21	formants	f5_max	0.939182
22	formants	f3_median	0.934186
23	formants	f4_max	0.839959
24	formants	f3_max	0.750493
26	harmonics-to-noise ratio	h2n_stdev	0.691567
27	formants	f1_median	0.597408
28	formants	f1_mean	0.553273
29	formants	f1_min	0.545544
30	long-term average spectrum	ltas_fmin	0.507023
31	formants	f2_stdev	0.498375
33	formants	f2_median	0.419127
34	formants	f2_max	0.409044
35	harmonics-to-noise ratio	h2n_diff	0.403711
36	pitch	f0_max	0.365185
37	long-term average spectrum	ltas_lph	0.361591
38	harmonics-to-noise ratio	h2n_max	0.351710
39	formants	f2_mean	0.349525
40	formants	f2_min	0.342969
41	formants	max_disp	0.281981
42	pitch	f0_stdev	0.225521
43	pitch	f0_diff	0.222980
44	formants	f1_max	0.200953
46	formants	f1_stdev	0.189932
47	energy	en_avg	0.162019
48	long-term average spectrum	ltas_fmax	0.086992
49	energy	en_stdev	0.066931
50	energy	en_mean	0.062653
51	energy	en_min	0.059033
52	energy	en_diff	0.056007

PRAAT [23] but incorporates various modifications making it faster and easier to integrate into the diarization system.

The features can be assigned to five different categories: pitch, energy, formants, harmonics-to-noise ratio, and long-term average spectrum. Pitch (the speaking fundamental frequency) is calculated using the so-called accurate autocorrelation method as described in [23], where experiments are presented showing that the method is more accurate, more noise resistant, and more robust than alternative methods, including traditional autocorrelation. The pitch calculation in PRAAT has been continuously improved. Our PRAAT-based feature extraction library is based on version 4.5.14.

a) Pitch Features: The default pitch as well as pitch range available to the speaker is influenced by the length and mass of the vocal folds in the larynx [24]. The main differences between speakers with respect to pitch is related to their gender and age. The vocal folds of postpubertal men are longer and thicker with a lower modal frequency compared to women and children. Therefore, their available range of frequencies is about 87–415 Hz, while for women it is 184–880 Hz. Although it is possible to achieve vibrational frequencies outside these ranges, this usually involves a degradation in the quality of vibration. Individual speakers vary in the range of frequencies they are capable of producing as well as the range of frequencies they actually use in everyday speech. Hence, pitch can be regarded as a capable speaker discriminant feature which has been confirmed in numerous speaker recognition studies some of which are itemized in Section II.

In this study, from the actual pitch track, where every frame is assigned to the most likely pitch value (see Fig. 3, 2), various long-range statistics were calculated. Note that nonspeech and unvoiced frames do not affect the results as pitch is marked as "undefined" in these cases.

- mean: The average value.
- median: The value of the 50th percentile. The median is generally less sensitive to outliers than the mean.
- min, max: Instead of the actual minimal and maximal values, the 5th and 95th percentiles, respectively, are taken to avoid an otherwise large impact of outliers caused by artifacts.
- diff: The difference between max and min as a measure of the local range.
- stdev: The standard deviation as a measure of the variance.
- swoj: The slope of the pitch curve ignoring octave jumps.

b) Energy Features: Compared to pitch, changes in loudness (or energy) are much less directly induced by anatomical characteristics. Rather than that, they are predominantly relevant to the marking of stress and to express emotions (which is also the case for pitch but to a smaller proportion). Still, energy features are considered as potentially speaker discriminant and therefore used in the candidate list of features for this study. However, they are not hypothesized as getting a high rank.

Energy features are based on an intensity contour with values in dB(SPL) i.e., dB relative to the human auditory threshold for 1 kHz, [see Fig. 3 (3)]. The following statistics have been calculated on the basis of this contour: min, max, diff (the difference between the former two), mean, and stdev.

c) Format Features: Formants are concentrations of acoustic energy around particular frequencies at roughly 1000-Hz intervals [see Fig. 3 (4)]. Formants occur around frequencies that correspond to the speaker-specific resonances of the vocal tract and are therefore suitable measures to help recognize the speaker. However, formants also depend on the phonetic content. They occur only in voiced segments, i.e., vowels [a], nasals [n], liquids [l], voiced fricatives [v], and voiced stops [b]. Moreover, in contrast to oral vowels, voiced consonants also exhibit anti-resonances in the vocal tract at one or more frequencies depending on the positions of the articulators which attenuate or eliminate formants at or near these frequencies. With nasal sounds, the division of the vocal tract into the nasal and oral branch also causes anti-resonances

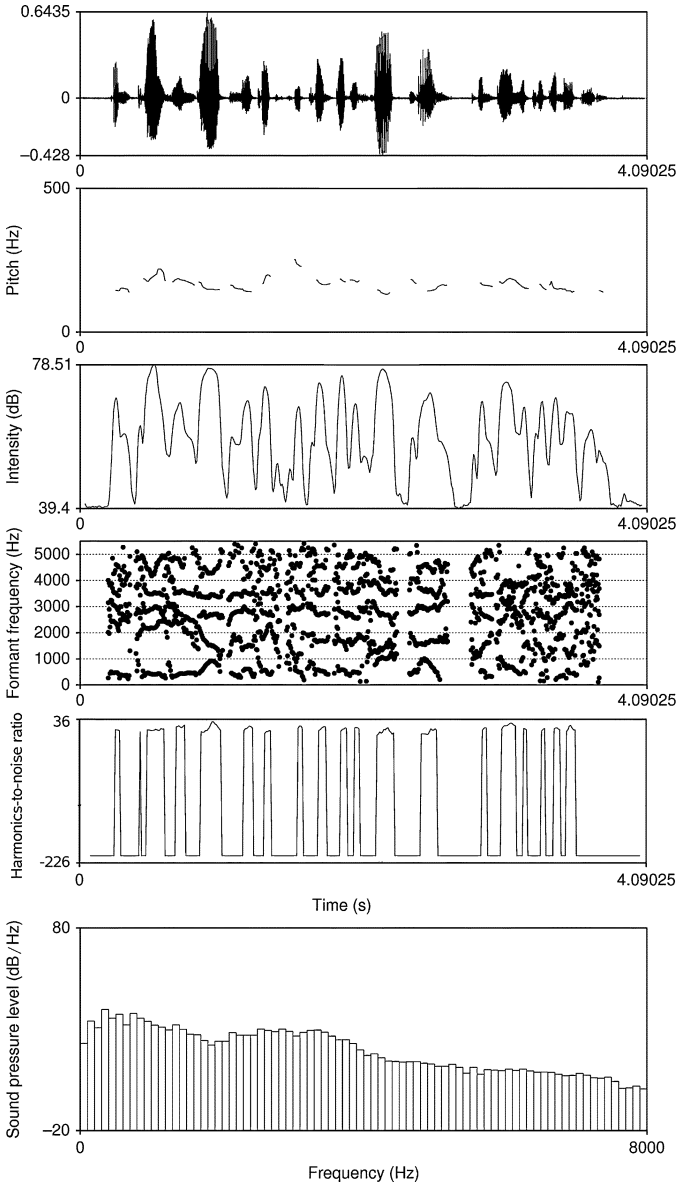


Fig. 3. Example waveform (1) along with: a pitch track (2), an intensity contour (3), a formant analysis (4), the harmonics-to-noise ratio (5), and long-term average spectrum (LTAS) (6).

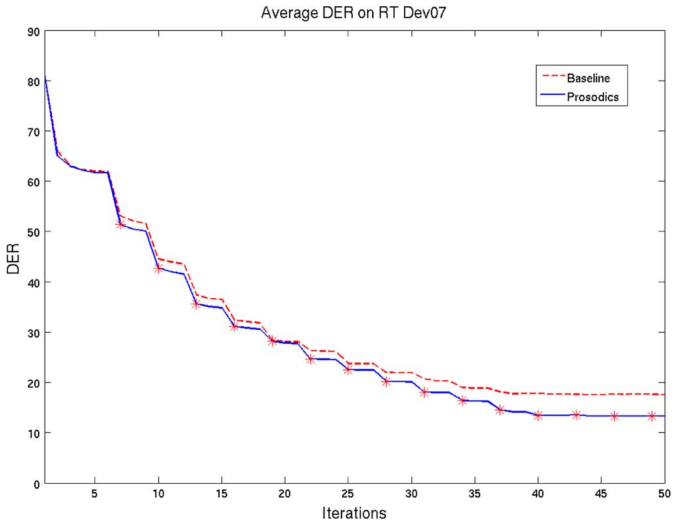


Fig. 4. Average DER per iteration of the ICSI Diarization Engine across the CombDev meetings (see text). There are 16 clusters and so the potential number of cluster merging is 15 (for meetings which the algorithm detects only one speaker). The asterisk denotes a merging of two clusters while other iterations are realignment of the models and the data. If, for one meeting, the algorithm stops at 40 iterations, the last value of the DER is used to compute the average for the last points of the graph.

while additional nasal formants are produced by resonances within the nasal branch. Even within one phonetic class, say oral vowels, the shape of the formants and the relation between them varies, which is obvious, as this is how the differences between the individual vowels are perceived. However, the variation related to the phonetic content happens for the most part in the first two formants while the higher ones are generally assumed to capture mainly speaker-specific information.

When measuring higher formants, bandwidth limitations have to be taken into account: the respective frequencies must be present in the signal, which is determined by the sampling frequency of the digitized signal. The sample rate of telephone speech is 8000 Hz, i.e., frequencies up to 4000 Hz are present (Nyquist limit). Hence, a maximum of three to four formants can be measured. The following formant-related statistics were used as candidate features (calculated for formants 1–5): mean, median, min (5th percentile), max (95th percentile), and standard deviation. Additionally, a formant dispersion measure was used. It was calculated as the sum of the differences (of min, max, and mean) between consecutive formants. The sampling rate of the NIST meeting data is 16 kHz so we consider frequencies up to 8 kHz.

d) Harmonics-to-Noise-Ratio Features: The harmonics-to-noise-ratio (HNR) quantifies the relative amount of additive noise in the voice signal. It is expressed in dB: if 99% of the energy of the signal is periodic and 1% is noise, then HNR $= 10 * \log 10(99/1) = 20$ dB. An HNR of 0 dB corresponds to an equal amount of harmonic and nonharmonic energy (see Fig. 3 (5)). Spectral noise can be caused by aperiodic vocal fold vibration and turbulent airflow generated by inadequate closure of the vocal folds during phonation [25]. The resulting friction noise is reflected in a higher noise level in the spectrum. HNR is therefore considered one of the parameters that can be used to quantify a perceptual impression of a rough voice. [26] reports that a healthy young speaker can produce a sustained [a] or [i] with a HNR of approximately 20 dB and a [u] with approximately 40 dB. The difference arises from the relatively high frequencies in [a] and [i] which make them more prone to fluctuations. This indicates also the limitations of using HNR for text-independent speaker recognition (or diarization): a fine-grained analysis requires the phone identity to be known. Also, a noisy channel compromises the usefulness of the feature. We nevertheless calculated mean, min, max, diff, and stdev. Undefined values (the low areas in the graph) which occur, for example, at silence frames, are ignored.

e) Long-Term Average Spectrum (LTAS): In order to obtain the long-term average spectrum, the spectral energy in 100-Hz-wide frequency bands is measured over a relatively large portion of speech (see Fig. 3). The standard deviation (stdev) was used as a measure of the variance. In addition, the

slope of the curve (slope), the frequency associated with the lowest energy (fmin) and highest energy (fmax), and the peak heights (lph) were calculated.

It would be computationally unfeasable to run diarization experiments with all subsets of this large set of prosodic features. We therefore decided to perform a computational inexpensive pre-experiment to select the most promising features. A visualization of the feature histograms suggested that most features roughly follow a Gaussian distribution. Therefore, from the abundant number of feature selection techniques [27], we decided to go with Fisher discrimant analysis (FDA). FDA provides a measure of the separatedness of classes, given that each class can be represented by a Gaussian.

To obtain a smaller set of features, we estimated their general speaker discriminability in a pre-experiment using the TIMIT database [28]. We are aware of the limitations of this database in terms of the lack of intersession and interchannel variability. However, these limitations are not crucial for the task of sub-selecting speaker-discriminant features [29]. TIMIT incorporates a large number of speakers (462) which are divided into roughly two-thirds male and one-third female, and each speaker has ten utterances. Also, the database is reasonably small, which reduced the complexity of the pre-experiment.

We measured the ratio of inter-speaker variance and intra-speaker variance. For each utterance, one value per feature was obtained (i.e., the range is the entire utterance). The relative speaker discriminability was estimated assuming that features perform better when the ratio between inter- and intra-speaker variance is higher. The *score* of a feature was therefore calculated as follows:

$$\text{score} = \frac{\Sigma_{i=1}\Sigma_{j=1}(\mu_i - \mu_j)(\mu_i - \mu_j)^T}{\Sigma_i \Sigma_{j:y_j=i}(x_j - \mu_i)^2}$$

where x represents a sample feature value, μ is the mean value for the feature for a given speaker i, or j, y_j is the speaker index for the jth sample.

We define that $\max(\text{score})$ indicates the best feature in the test. Basically, a feature with a high score is one where the class-means are well separated, measured relative to the (sum of the) variances of the data assigned to a particular class. It implies that the gap between the classes is expected to be big. The order of features on Table I reflects the results of the experiments (see the right-most column). The ten features with the highest score are selected.

Generally, the results appeal to our intuition: The median and average fundamental frequency are the best features, followed by high formants (F4, F5). Also, the mean harmonics-to-noise ratio and the variance of the long-term average spectrum achieved a high score. Although pitch_median and pitch_mean are likely to be highly correlated, we decided to keep them both since their scores are outstanding.

The validity of the approach is reflected in the final diarization experiments described in the next section.

V. Integration Into ICSI Speaker Diarization

Although it can be assumed that general speaker discriminability is important for speaker diarization, the performance of the selected features using a concrete engine was still to be evaluated. Therefore, we performed a second set of experiments using the ICSI Speaker Diarization system. In these experiments, the speech/nonspeech detection remains the same—only the feature input to the agglomerative clustering step (see Section III) is modified. All experiments were performed on the NIST RT07 evaluation data for single distant microphone condition. It contains eight meetings recorded in several geographic locations with differing numbers of people. This set is hereafter referred to as NIST Eval07. Even though the diarization task is unsupervised, some parameters can be adjusted and optimized, such as the initial number of Gaussians or the weights for the various feature streams. Another set of 21 meetings, based on all NIST RT evaluation and development meeting data of the previous years (excluding the evaluation data Eval07), is used for parameter selection (hereafter referred to as CombDev). Basic statistics of the data sets are shown in Table II.

TABLE II
Statistics of the NIST RT Meeting Evaluation Data. Min/Max Value for the Meeting Length (in Seconds) and the Number of Speakers, the Number of Speaker Turns, and the Average Duration Per Speaker Turn (in Seconds)

Data set	Length (s)	#speakers	#turns	Avg. turn length
CombDev	693 / 3946	2 / 6	8646	3.68s
Eval07	1352 / 2826	2 / 6	5424	3.04s

In the first experiment, we ran the diarization engine using the ten most speaker-discriminant features based on the selection study described in Section IV. The results are far from being competitive with the state of the art (ICSI baseline system). Basically, the segments obtained at the end were, in most cases, clusters correlating to gender. Finer voice characteristics could not be discriminated. This comes to no surprise as prosodic features capture only longer term characteristics of the audio signal. However, having performed the experiments described in Section IV and the fact that short term features are successfully being used for diarization, suggest that there is a potential improvement of DER by using long-term and short-term features in combination.

In order to combine the new set of features with "traditional" MFCCs, one feature value per frame must be extracted while maintaining a minimum length of the actual extraction region. This is obtained by using a Hamming window of 500 ms and a step size of 10 ms. Another issue is how to deal with missing values. Pitch features, for example, are naturally undefined on unvoiced regions of speech. We applied the most straightforward solution of replacing the undefined values by the mean value of the respective feature calculated over the entire meeting.

The approach we propose for combining several features is similar to the one in [30]. In particular, the function performed by the diarization engine is to maximize the likelihood of the observed data given the model (in our case, the model is an ergodic HMM). We can then define the combined likelihood for the emission probabilities as

$$p(x_{\text{MFCC}}, x_{\text{PROS}} \mid \theta_i) = p(x_{\text{MFCC}} \mid \theta_{i1})^{1-\alpha} p(x_{\text{PROS}} \mid \theta_{i2})^{\alpha}$$

TABLE III
DER BREAKDOWN FOR THE COMBDEV (SEE TEXT) DATA BY USING MFCC+PROSODIC FEATURES (BASELINE IS MFCC ONLY). Sp/nsp (SPEECH/NONSPEECH) IS THE ERROR DUE TO SPEECH ACTIVITY DETECTION (SAME SYSTEM AS IN BASELINE) WHILE SpkrSeg IS THE ERROR DUE TO THE SPEAKER SEGMENTATION ALGORITHM

Meeting ID	Sp/nsp	SpkrSeg	Total DER
AMI_20041210-1052	0.90 %	0.60 %	1.50 %
AMI_20050204-1206	3.30 %	3.70 %	7.01 %
CMU_20050228-1615	8.30 %	1.60 %	9.90 %
CMU_20050301-1415	3.40 %	1.80 %	5.19 %
CMU_20050912-0900	14.80 %	8.20 %	22.97 %
CMU_20050914-0900	12.90 %	2.30 %	15.20 %
EDI_20050216-1051	4.60 %	18.20 %	22.79 %
EDI_20050218-0900	5.80 %	11.70 %	17.43 %
ICSI_20000807-1000	4.80 %	3.80 %	8.55 %
ICSI_20010208-1430	4.80 %	10.60 %	15.36 %
LDC_20011116-1400	4.90 %	0.90 %	5.78 %
LDC_20011116-1500	7.60 %	6.60 %	14.13 %
NIST_20030623-1409	1.60 %	5.70 %	7.25 %
NIST_20030925-1517	11.30 %	5.70 %	17.08 %
NIST_20051024-0930	6.40 %	1.00 %	7.41 %
NIST_20051102-1323	5.30 %	17.90 %	23.27 %
TNO_20041103-1130	7.00 %	11.30 %	18.32 %
VT_20050304-1300	1.60 %	2.80 %	4.42 %
VT_20050318-1430	3.50 %	16.70 %	20.24 %
VT_20050623-1400	8.20 %	5.40 %	13.57 %
VT_20051027-1400	7.80 %	7.90 %	15.64 %
ALL	6.20 %	7.00 %	13.29 %
ALL (baseline)	6.40 %	11.30 %	17.57 %

TABLE IV
DER BREAKDOWN FOR THE NIST EVAL07 DATA BY USING MFCC+PROSODIC FEATURES (BASELINE IS MFCC ONLY). Sp/nsp (SPEECH/NON-SPEECH) IS THE ERROR DUE TO SPEECH ACTIVITY DETECTION (SAME SYSTEM AS IN BASELINE) WHILE SpkrSeg IS THE ERROR DUE TO THE SPEAKER SEGMENTATION ALGORITHM

Meeting ID	Sp/nsp	SpkrSeg	Total DER
CMU_20061115-1030	13.9 %	9.1 %	22.98 %
CMU_20061115-1530	6.7 %	8.6 %	15.25 %
EDI_20061113-1500	10 %	16.5 %	26.43 %
EDI_20061114-1500	6.2 %	14.6 %	20.75 %
NIST_20051104-1515	3.8 %	1.5 %	5.29 %
NIST_20060216-1347	3.3 %	4.1 %	7.43 %
VT_20050408-1500	5 %	2 %	7.05 %
VT_20050425-1000	5.7 %	21.6 %	27.36 %
ALL	6.80 %	9.50 %	16.28 %
ALL (baseline)	6.80 %	15 %	21.81 %

where x_{MFCC} and x_{PROS} represent the feature vectors (the MFCC vector being 19-dimensional and the prosodic vector being ten-dimensional), θ_{i1} represent the parameters of cluster i using the MFCC features extracted from the observed data and θ_{i2} are the parameters using the prosodic features. The model we use for the emission probabilities are GMMs where the number of components varies for each feature stream. Note that there is an assumption of independence between the two sets of features. Finally, as we observed that MFCC features alone tend to perform better than prosodic features, we used the α parameter to weight the confidence given to each feature stream. If α is set such that $\alpha < 1$, the likelihoods of the prosodic features given each class are flattened (the extreme case where $\alpha = 0$ map all the likelihoods to 1). Hence, the effect of this parameter is to give a different confidence value to each feature stream.

The CombDev set is used to find the optimal value for α. The initial number of Gaussians of the prosodic features is set to 2 and we use the top ten performing prosodic features. The rest of the parameters are the same as the ones used in the RT07 evaluation (16 initial clusters and five Gaussians per cluster for the MFCC feature vector).

Table III shows the results on the development set with the optimal value of $\alpha = 0.1$. The use of the top-ten prosodic features resulted in a 24.36% relative improvement of the DER (from 17.57% to 13.29% DER absolute). Table IV shows the results using the top ten features on the Eval07 set, compared to the system that performed best in the NIST Evaluation for the SDM condition (baseline system). The relative improvement is 25.36%, which is consistent with what was observed on the CombDev data.

Fig. 4 shows the DER evolution per algorithm stage of the baseline system versus our combined approach. As can be seen, the top ten prosodic features contribute especially in the last stages of the agglomerative clustering approach. Since the α value found using the development set was low, the effect of the prosodic features on the first iterations is unnoticed by the algorithm: the MFCCs alone are able to refine the segments and merge clusters that belong to the same speaker. As the clusters are merged, the average length increases and thus the long-term dependencies that the prosodic features extract are more robust. Moreover, in the last stages of the algorithm the clusters are more pure (each cluster contains speech from only one person), and, as a consequence, the discriminative power that the prosodic features have is amplified by the fact that the clusters represent speech from mostly one person. If we observe the tail of Fig. 4, it is clear that the information provided by the prosodic features is quite useful in the last stages, where the MFCCs would not otherwise have been able to correct some errors.

We also conducted diarization experiments using the top 11–20 ranked prosodic features on CombDev. The resulting DER is 17.29% absolute, which is almost the same as the baseline system. Compared to the 24.36% relative improvement generated from using the top ten ranked features, the use of the top 11–20 ranked features does not give significant improvement over the baseline system, which further verifies our feature selection approach discussed in Section IV.

VI. CONCLUSION AND FUTURE WORK

In this paper, we showed how a selection of prosodic and other long-term features in combination with MFCCs dramatically increases the accuracy of a state-of-the-art speaker diarization system. In combination with the commonly used MFCCs, we observed a significant improvement of our system with respect to the official submission on the last NIST RT 2007 evaluation task. This result was also verified on a wide set of meetings, which we call CombDev, that contains 21 meetings from previous evaluations. Since the prosodic and long-term features were selected using a diarization-independent speaker-discriminability study, we are confident that the same features are able to improve other systems that perform similar tasks.

Future work includes making the calculation of prosodic features more computationally efficient. Currently, the computation of a set of ten prosodic features takes about $1 \times$ real-time on a 2.2-GHz Athlon PC. This means, when adding prosodic

features, the computation effort of the entire diarization process takes almost twice as long as the duration of the audio file.

We will try "pyramid approaches" where multiple layers of feature streams with different time resolutions are used together. Similar approaches are used in image processing as well as in speech processing for a large range of applications. We believe that investigating different front-end features is equally important to creating new machine learning approaches. The use of alternative front-end features improves the robustness of speaker diarization and provides further insights for understanding the underlying basic research problem of speaker discrimination.

ACKNOWLEDGMENT

The authors would like to thank N. Morgan, A. Janin, K. Boakye, C. Wooters, N. Mirghafori, and M. Huibregts for their input. The authors would also like to thank M. Feld for creating an easy-to-use C-interface to access the prosodic feature computation functionality inside PRAAT.

REFERENCES

[1] D. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'05)*, Mar. 2005, vol. 5, pp. 953–956.

[2] E. Shriberg, "Higher-Level Features in Speaker Recognition," in *Speaker Classification I*, ser. Lecture Notes in Artificial Intelligence, C. Müller, Ed. Heidelberg, Germany: Springer, 2007, vol. 4343.

[3] K. Soenmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *Proc. Int. Conf. Spoken Lang. Process. 1998*, 1998, no. 0920.

[4] A. Janin *et al.*, "The ICSI meeting corpus," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03)*, 2003, vol. 1, pp. 364–367.

[5] S. Kajarekar, L. Ferrer, K. Soenmez, J. Zheng, E. Shriberg, and A. Stolcke, "Modeling NERFs for speaker recognition," in *Proc. Speaker Odyssey*, 2004, pp. 51–56.

[6] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, and D. Reynolds, "Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS'02," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03)*, 2003, vol. 4, pp. 729–795.

[7] D. Reynolds *et al.*, "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03)*, 2003, vol. 4, pp. 784–787.

[8] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," in *Speech Commun.*, Jul. 2005, vol. 46, no. 3–4, pp. 455–472.

[9] L. Ferrer, E. Shriberg, S. Kajarekar, and K. Sonmez, "Parameterization of prosodic feature distributions for SVM modeling in speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., (ICASSP'07)*, 2007, vol. 4, pp. 233–236.

[10] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. DARPA Speech Recognition Workshop*, 1998.

[11] H. Ning, M. Liu, H. Tang, and T. Huang, "A spectral clustering approach to speaker diarization," in *Proc. Interspeech*, 2006, ISCA, article ID: 1607-ThuA10.1.

[12] B. H. Juang and L. R. Rabiner, "A probabilistic distance measure for hidden Markov models," *AT&T Tech. J.*, vol. 64, no. 2, pp. 391–408, 1985.

[13] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system," in *Proc. NIST MLMI Meeting Recognition Workshop*, Edinburgh, U.K., 2005, pp. 402–414, Springer.

[14] P. Delacourt and C. Wellekens, "Distbic: A speaker-based segmentation for audio data indexing," *Speech Commun.: Special Iss. Access. Inf. Spoken Audio*, vol. 32, no. 1–2, pp. 111–126, 2000.

[15] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Process. Mag.*, vol. 11, no. 4, pp. 18–32, Oct. 1994.

[16] J. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Oct. 1997.

[17] H. Kim, D. Ertelt, and T. Sikora, "Hybrid speaker-based segmentation system using model-level clustering," in *IEEE Int. Conf. Acoust., Speech, Signal Process., (ICASSP'05)*, 2005, vol. 1, pp. 745–748.

[18] A. Gallardo-Antolin, X. Anguera, and C. Wooters, "Multi-stream speaker diarization systems for the meetings domain," in *Proc. Interspeech*, 2006, no. 1620-Thu1A1O.3.

[19] J. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences," in *Proc. Interspeech*, 2006, no. 1337-Thu1A1O.5.

[20] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Proc. NIST RT07 Meeting Recognition Evaluation Workshop*, 2007, pp. 509–519, Springer.

[21] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *2003 IEEE Workshop Autom. Speech Recognition Understanding ASRU'03*, 2003, pp. 411–416.

[22] Y. Huang, O. Vinyals, G. Friedland, C. Müller, N. Mirghafori, and C. Wooters, "A fast-match approach for robust, faster than real-time speaker diarization," in *Proc. IEEE Autom. Speech Recognition Understanding Workshop*, 2007, pp. 693–698.

[23] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. Dutch Inst. Phon. Sci. (IFA)*, Amsterdam, The Netherlands, 1993, pp. 97–110.

[24] V. Dellwo, M. Huckvale, and M. Ashby, "How is individuality expressed in voice? An introduction to speech production & description for speaker classification," in *Speaker Classification*, ser. Lecture Notes in Computer Science/Artificial Intelligence, C. Müller, Ed. New York: Springer, 2007, vol. 4343.

[25] C. Ferrand, "Harmonics-to-noise ratio: An index of vocal aging," *J. Voice*, vol. 16, no. 4, pp. 480–487, 2002.

[26] P. Boersma, "PRAAT, a system for doing phonetics by computer," *Glot Int.*, vol. 9, no. 5, pp. 341–345, 2001.

[27] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[28] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet, and N. L. Dahlgren, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM," National Inst. Standards Technol., Gaithersburg, MD, Tech. Rep. NISTIR 4930, 1993.

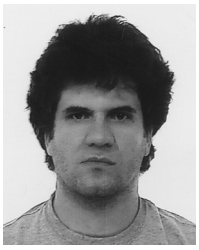
[29] C. Mueller, *Sprecherklassifikation nach Alter und Geschlecht*. Heidelberg, Germany: Akademische Verlagsgesellschaft Aka GmbH, 2006.

[30] A. Gallardo-Antolin, X. Anguera, and C. Wooters, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Trans. Comput.*, vol. 56, no. 9, pp. 1212–1224, Sep. 2007.

Gerald Friedland (M'08) received the "diploma" and doctorate (with distinction) in computer science from Freie Universität Berlin, Berlin, Germany, in 2002 and 2006, respectively.

He is a Staff Scientist with the International Computer Science Institute (ICSI), an independent nonprofit research lab associated with the University of California at Berkeley. He is currently leading the Speaker Diarization research. Apart from speech processing, his interests also include computer vision and multimodal machine learning.

Dr. Friedland is involved in the organization of various ACM and IEEE conferences, including the IEEE International Conference on Semantic Computing (ICSC2009), where he serves as co-chair and the IEEE International Symposium on Multimedia (ISM2009) where he serves as program co-chair. He is the recipient of several research and industry recognitions, among them the Multimedia Entrepreneur Award by the German government and the European Academic Software Award.



Oriol Vinyals received the double master's degree mathematics and telecommunications from the Polytechnic University of Catalonia, Barcelona, Spain. He is currently a graduate student in computer science at the University of California, San Diego.

Before doing an internship with Microsoft Research, he was a Visiting Scholar at the International Computer Science Institute (ICSI), University of California, Berkeley, for more than a year where he had worked on topics which involve speaker diarization, speaker recognition, and machine learning. This work was conducted while he was with ICSI.

Yan Huang (M'08) received the M.S.E. degree in electrical engineering from The Johns Hopkins University, Baltimore, MD, in 2001, and the M.S. degree in computer science from the University of California, Berkeley, in 2007.

Previously, she has been with International Computer Science Institute (ICSI), Berkeley, CA, Panasonic Speech Technologies Laboratory, Santa Barbara, CA, and the Center of Language and Speech Processing, Baltimore, MD. She has been working on various components of large vocabulary speech recognition systems, speaker diarization, speaker recognition, and speech synthesis. She is currently a Research Scientist with Li Creative Technologies, Inc., Florham Park, NJ. Her major research interest is in machine learning and its applications in speech and language processing. This work was conducted while she was with the International Computer Science Institute (ICSI).

Christian Müller studied computational linguistics at Saarland University, Saarbrücken, Germany, where he graduated with distinction in 1994 and received the Ph.D. degree in computer science at Saarland University in January 2006. His dissertation "Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht" (Two-layered Context-Sensitive Speaker Classification on the Example of Age and Gender) was supervised by Prof. W. Wahlster.

He possesses more than a decade of professional experience in the field of speech technology. Currently, Christian Müller is a Senior Researcher at the Germany Research Center for Artificial Intelligence (DKFI), Saarbrücken. From 2006 to 2008, he was a Visiting Researcher at the International Computer Science Institute (ICSI), Berkeley, CA. This work was conducted while he was with ICSI.