Unfolding Speaker Clustering Potential: A Biomimetic Approach

Thilo Stadelmann Bernd Freisleben
Department of Mathematics & Computer Science, University of Marburg
Hans-Meerwein-Str. 3, D-35032 Marburg, Germany
{stadelmann, freisleb}@informatik.uni-marburg.de

ABSTRACT

Speaker clustering is the task of grouping a set of speech utterances into speaker-specific classes. The basic techniques for solving this task are similar to those used for speaker verification and identification. The hypothesis of this paper is that the techniques originally developed for speaker verification and identification are not sufficiently discriminative for speaker clustering. However, the processing chain for speaker clustering is quite large – there are many potential areas for improvement. The question is: *where* should improvements be made to improve the *final* result? To answer this question, this paper takes a biomimetic approach based on a study with human participants acting as an automatic speaker clustering system. Our findings are twofold: it is the stage of modeling that has the highest potential, and information with respect to the temporal succession of frames is crucially missing. Experimental results with our implementation of a speaker clustering system incorporating our findings and applying it on TIMIT data show the validity of our approach.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing; I.5.4 [**Pattern Recognition**]: Applications—*Signal processing, Waveform analysis*

General Terms

Algorithms, Design, Experimentation, Performance

Keywords

Speaker identification, Speaker clustering, Speaker diarization, GMM, MFCC, Temporal context, One-class SVM

1. INTRODUCTION

Recognizing voices automatically is useful for several applications. For example, it supports biometric authentication [64]. It helps making speech recognition robust [20]. It enables search engines to index spoken documents and thus improves retrieval performance [31]. These three examples refer to different subproblems of speaker recognition, namely: speaker verification [49], speaker identification [8] and speaker clustering [28] (or, when regarding the complete process including speech detection and segmentation: speaker diarization [46]).

Speaker verification is the most simple clustering task among these problems: the question is whether a given utterance can be assigned to a given model (identity) – a binary choice. Speaker identification is a $(1 : n{+}1)$ choice: the question is which (if any) of the given models can the given utterance be paired with? Finally, speaker clustering is a $(m : n)$ problem in which all utterances are equally important and each utterance may be grouped together with any other utterance – or stay alone. Both the number of clusters (speakers) and the actual cluster memberships must be determined automatically.

The speaker verification and identification tasks have been studied extensively in the literature. Using Mel Frequency Cepstral Coefficients (MFCCs) [12] as parametric speech features and Gaussian Mixture Models (GMMs) [49] (with more recent modifications [48]) as speaker models has become the quasi-standard, although other methods have been proposed [16]. This is due to quite satisfactory results with just moderate demands for the data: the utterances should be relatively noise-free (telephone speech works) and long enough (minimum 10 seconds, better more than 30 seconds per utterance) [62]. The canonical example is the experiment in Reynolds' classic paper on GMMs [47]: The 630 speakers of the TIMIT database [19] are split into a training set (8 sentences per speaker concatenated to one utterance) and a separate test set (2 sentences per speaker form one utterance). Each sentence is approximately 3 seconds long. The utterances are transformed to MFCC feature vectors. For the 630 training utterances, GMMs with 32 mixtures are built a priori, then an identification experiment is run for the 630 test utterances. It yields a satisfactory 0.5% closed set identification error.

Speaker clustering has also been studied extensively for more than a decade [24]. The basic techniques used for speaker clustering are largely along the lines of the previously discussed verification/identification techniques: MFCC features are modeled by GMMs [28][60]. Upon this, a step-by-step scheme using agglomerative hierarchical clustering is usually built using some metric (often the Generalized Likelihood Ratio (GLR)) and a termination criterion (frequently based on the Bayesian Information Criterion (BIC))

MM'09, October 19–24, 2009, Beijing, China.

[34]. Evaluations typically concentrate on data sets built from broadcast news/shows and meeting recordings, where diarization error rates ranging from 8% to 24% are reported [28][34][45]. These results are confirmed by more recent approaches that otherwise deviate from the standard methodical scheme (e.g. by using genetic algorithms instead of agglomerative clustering [61] or Support Vector Machines (SVMs) instead of GMMs [17]).

From the definition of the task of speaker clustering it is evident that speaker clustering has a much higher complexity than the other two tasks. This fact certainly affects the anticipated outcome in terms of higher expected error rates and/or applicability only to less complex data. Both implications can be observed in the literature:

1. Error rates for clustering and identification are significantly apart from each other, as indicated above.

2. Data sets for clustering have a considerably smaller speaker population size: for example, in the approaches surveyed by Kotti et al. [28], the number of speakers (with several segments each) per run ranges from 2 to 89, with an average of 28 speakers (and a standard deviation of 31) as compared to 630 in the speaker identification example above. As pointed out by Reynolds [49], a smaller number of speakers eases the task considerably.

3. Several authors notice that the current clustering or diarization systems are not very robust to data variation and thus are poorly portable [46][23][66]. This is in contrast to the wide applicability of speaker verification and identification techniques [42][6].

In this paper, we present an experiment to determine what impact the change in a experimental setting (i.e., from identification to clustering) has on the results. We used the basic settings of Reynolds' identification experiment on TIMIT [47] and re-ran it with our own implementation of the complete speaker identification chain. It yielded 0.0% closed set identification error (we attribute the difference to Reynolds' original results to subtle varieties in the implementations of the signal processing and model initialization parts). We then changed the experimental setting from an identification scenario to clustering (i.e., each of the 1260 utterances can now be grouped with any other utterance; before, there was prior knowledge that 630 utterances are distinct speakers and each of the remaining 630 utterances has to be grouped with an utterance of the first group). Our speaker clustering software uses the same framework as the identification module and implements a state-of-the-art system comparable to the one described by Han et al. [23] (of course without the "selective clustering" part that would nearly reduce our clustering experiment to the identification task for optimal parameter settings).

The system scored a misclassification rate of 99.84% with respect to utterances, which effectively shows that the task is too complex for the used techniques. In contrast to the identification task before, efforts were made to find optimal parameter settings for the values that did not correspond to settings in Reynolds' experiment and thus should not be altered for the sake of comparability. For 16 kHz data (Reynolds used 8 kHz), we used: MFCCs 1–19 (coefficient 0 discarded) extracted from 20 ms long frames every 10 ms using a 512 point Fast-Fourier Transform (FFT) on the Hamming-windowed, pre-emphasized ($\alpha = 0.97$) signal and a mel filterbank of 24 triangular filters ranging from 0 to 7600 Hz. GMMs with 32 mixtures and diagonal covariances were initialized via a maximum of 10 iterations of k-means seeded by the deterministic Var-Part method [56] and trained with a maximum of 15 Expectation-Maximization (EM) steps (or until the increase in likelihood dropped below 100, whatever happened first) having a variance limit of 0.01. Individual models were compared using the distance measure described by Beigi et al. [5] (in conjunction with the Euclidean distance between single mixtures). Clustering was performed based on these distances using complete linkage and stopped by the Information Change Rate (ICR) measure tuned to the optimal threshold using ground truth data. The choice of the metric, linkage method and termination criterion was motivated by comprehensive experiments comparing most of all reasonable options and choosing the best for this task on a subset of the data.

The encountered complexity is distinct (in fact: additive) in nature to what is described by Morris et al. [39] to make identifying voices on TIMIT data a challenge: the pure quantity of speakers seems to exhaust the expressive power of the clustering system in the presence of an increased number of degrees of freedom. This view is supported by the fact that the same clustering experiment performs relatively well (12.50% misclassification rate) for a reduced subset of only the first 40 speakers out of the original 630 and even perfect for 20 speakers and less.

The hypothesis of this paper is: the techniques originally developed for speaker verification and identification are not suitable for speaker clustering, taking into account the escalated difficulty of the latter task. However, the processing chain for speaker clustering is quite large – there are many potential areas for improvement. The question is: *where* should improvements be made to improve the *final* result?

In this paper, first we show which part of the processing chain bears how much potential for further improvement. This part of the answer implies that improving other parts of the chain will probably not show the full potential of that improvement: an improvement at the beginning of the pattern recognition process is probably not able to propagate until its end if it is succeeded by an even greater source of failure. Second, we state explicitly what this improvement has to look like qualitatively. Third, we present an implementation of a speaker clustering system that experimentally supports our thesis by improving existing results on a TIMIT benchmark test. Our approach is based on an analysis of the operating mode and capability of the best speaker clustering automaton available: the human being, according to the principle of biomimetics [4].

The paper is organized as follows. The design of a speaker grouping study with human participants is described in Section 2. The evaluation and interpretation of the results of the study follows in Section 3. Section 4 presents a technical implementation of our findings in a speaker clustering systems along with corresponding results. Section 5 concludes the paper and outlines areas for future research.

2. ANALYZING THE PROCESS

This section reports on the motivation, design, technical background and results of a study that puts humans in the role of a speaker clustering software: participants are asked to group together utterances based on their inferred speaker

identity within variants of the same data set. These variants are the internal representations of the original speech signal at different levels of the pattern recognition process in an actual speaker clustering software made audible again.

2.1 Motivation

The poor results of the speaker clustering experiment on the full TIMIT database raise the question what kind of information is actually missing in the applied methods. The feature extraction method at the beginning of the pattern recognition chain lossily compresses the information included in the original signal [7], and the later speaker modeling (i.e., classifier training) stage basically does the same. The basic idea of our approach is to use the qualitative judgment of humans based on their experience as listeners to determine the lacking information in the different pattern recognition stages. This requires to represent the acoustic signal at these stages such that the participants can listen to it, i.e., resynthesis. From the evoked sensation, the level of discernability present in the data is determined: signals sounding very similar might also be difficult to distinguish by a computer. This is measured by asking our participants to perform a speaker clustering experiment that is evaluated in the same way a software system would be evaluated.

The rationale is: we already have demonstrated above that the clustering software succeeds for a reduced TIMIT data set of less than 40 speakers. If humans find a reasonable clustering for the original speech signal but cannot distinguish the data as used by the computer – showing that the computer essentially does not have some information that was still present in the original signal – there is some unused potential. This potential lies in the information that was removed in the course of processing.

Several arguments support our approach: Humans may not be trained to analyze synthetic speech features, but in contrast to machine learning techniques that need well-posed learning problems [37] as well as an appropriate training data basis – human learning is generalizing well and adaptive [21]. Information is best (i.e., very quickly and reasonably accurately) grasped with our auditory system as a guide in an otherwise unstructured search in a large hypothesis space [13]. A similar view has been advertised by Aucouturier in the field of music information retrieval [2].

2.2 Design

The primary goal of our study is to show which stage of the processing chain of speaker clustering bears how much potential for improvement (then, what can and has to be improved). The two stages of feature extraction and modeling are the most promising candidates, since there the main information reduction takes place. Further candidate stages are signal (pre-)processing (which we add to feature extraction), segmentation (into e.g. silence/speech/noise, which are complete pattern recognition processes in themselves and therefore are likely to benefit from this study rather than contribute to it) and clustering (which is not considered here for reasons explained later in Section 3). To accomplish our goal, we apply the biomimetic approach of observing human behavior. To obtain relevant results, we have created a feasible data set along with a test philosophy and have acquired a reasonable group of participants.

The data set is based on a subset of the TIMIT data mentioned in Section 1 with a meaningful but manageable size. It contains 7 speakers, hence 14 utterances, with 3 male and 4 female voices from the same dialect region. We took the first 7 speakers in lexicographical ordering of the file names: FAKS0, FDAC1, FELC0, FJEM0, MDAB0, MJSW0 and MREB0 from TEST/DR1. The data set (and additional material for reproducing the study) is publicly available at `http://www.informatik.uni-marburg.de/~stadelmann/download/sg_experiment.zip`. Reynolds' procedure is used to concatenate the 10 sentences to 2 utterances per speaker (see Section 1). This material is the input to our speaker clustering system, scoring perfectly with 0.0% error. As side products, the system outputs altered versions of the input data (equal to it in length), namely resynthesized features and resynthesized models (the technical details of this process are presented in the next subsection). This yields "dataset 1" (resynthesized speaker models, sounding similar to "bubbling/boiling liquid"), "dataset 2" (resynthesized feature vectors, sounding like a "robot voice") and "dataset 3" (original speech, sounding "normal") for the human speaker grouping study.

According to our test plan, the three data sets are presented to the participants in the order described above. The task is the same for each data set: within 30 minutes or less (to set an upper bound on the time for participation), a participant is supposed to group the 14 utterances by the inferred speaker identity. This is done by drawing lines between the utterances in question on the assessment sheet, where their file names (i.e., numbers) are arranged on a circle. The participants are told to "engage" with the sound and "not to focus on maybe unfamiliar patterns that all recordings of a run have in common, but on the more subtle differences, like the ones used when, for example, distinguishing two low-pitched male voices. The decision to group recordings together must be taken solely based on the acoustical similarity of the voices". By hearing the more unfamiliar sounds first, it is ensured that no participant is tempted to transfer findings from an earlier data set to a later one. To further minimize such effects, the arrangement of the utterances on the assessment sheet's circle is permuted randomly between runs. Together with the actual grouping, the participants are asked to describe "in 1–3 short sentences how [they] tried to solve the task and how [they judged their] own result". The freedom offered by this formulation is intentional so that driving the participants in any direction by asking specific questions on used features, methods or experienced difficulties is prohibited. These instructions are given to the participants together with the data.

Our group of participants consists mainly of students and university staff ranging in age from 21 to 64 years (mean: 30.7, standard deviation: 8.98). Overall, 20 people participated, 6 of them being female and 14 male, giving a representative sample in size and composition. Each participant is told to read the instructions and act accordingly. This effectively eliminates prior knowledge on the design and goal of the study. The comprehensibility and sufficiency of the instructions and the feasibility of the task has been approved in a pretest.

2.3 Technical Background

Although we report on a study with human participants, a technical challenge to deal with is the reconversion of features and models to speech. The design of the corresponding tools is presented in this subsection.

Table 1: Comparison of human and random clustering using statistical measures

means	dataset	time [m]	#clusters	#correct	#connections	FAKS0	FDAC1	FELC0	FJEM0	MDAB0	MJSW0	MREB0
human $\mu(\sigma)$	1	22.95 (7.44)	6.05 (2.39)	3.0 (1.72)	8.05 (2.52)	0.25	0.4	0.55	0.45	0.4	0.25	0.7
random $\mu(\sigma)$	1	-	6.49 (1.48)	1.04 (1.05)	7.51 (1.48)	0.14	0.15	0.15	0.15	0.14	0.14	0.14
human wins?	1	-	no	0.0005	0.1	0.1	0.001	0.0005	0.0005	0.0005	0.1	0.0005
human $\mu(\sigma)$	2	17.33 (7.71)	6.35 (1.31)	3.3 (1.92)	7.75 (1.41)	0.25	0.6	0.7	0.35	0.4	0.4	0.6
random $\mu(\sigma)$	2	-	6.77 (1.23)	0.85 (0.91)	7.23 (1.23)	0.12	0.12	0.13	0.13	0.13	0.12	0.11
human wins?	2	-	no	0.0005	0.1	0.05	0.0005	0.0005	0.005	0.0005	0.0005	0.0005
human $\mu(\sigma)$	3	8.95 (5.19)	7.2 (0.77)	6.55 (1.05)	6.75 (0.72)	0.85	1.0	0.95	0.85	0.95	0.95	1.0
random $\mu(\sigma)$	3	-	7.37 (0.57)	0.51 (0.72)	6.63 (0.57)	0.07	0.08	0.08	0.08	0.07	0.07	0.07
human wins?	3	-	no	0.0005	no	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005

At the beginning, the concrete realization of "features" and "models" must be defined. MFCCs are used as features and GMMs as representatives of speaker models, based on the reasoning explained in Section 1: MFCCs are by far the most popular features, and GMMs with diagonal covariances are very often used as models. Preliminary listening experiments among different types of models did not reveal substantial audible differences for the resynthesized voices. Tested modeling techniques include the multivariate full-covariance Gaussian model, GMM with diagonal- and full covariances, and a left-to-right Hidden Markov Model (HMM) with 5 states and 5 mixtures per state. Furthermore, this study's aim is to refer to the experiment conducted by Reynolds [47], in which GMMs modeling MFCCs were used.

The primary requirement on the resynthesized voices is to make audible what is contained in the models and feature vectors. Thus, no effort is made to make the result more intelligible or natural beyond what this data originally contains and conveys.

Next, the inversion process from a model back to an audio file is described. A GMM is a statistical model that represents a probability distribution. Feature vectors following the distribution can be obtained via sampling from the GMM. This is a two-stage process: First, a mixture component i_m is chosen at random according to the distribution determined by the mixture weights. This is accomplished by generating a uniformly distributed random number r in $[0, 1]$ and then summing up the mixture weights until the sum exceeds r; the mixture index i_m of the last added weight (of course, weights are ordered in the same way each time) subscripts the chosen mixture component. Second, a normal deviate is drawn from mixture component i_m via the polar (Box-Muller) method [26]. Because the GMM was built from MFCC feature vectors, the resulting random vector is also a valid MFCC vector.

Converting MFCCs back to a waveform includes the following steps: the effects of the preemphasis filter and the mel filterbank need to be canceled out, preferably in the cepstral domain (there it reduces to subtracting these two filters' cepstra). The circumsized cepstrum is then filled up with zeros and transformed back to the log filterbank domain by the inverse DCT, where the log() operation is remedied and the filterbank is reversed via an overlap-and-add method. This yields a standard magnitude spectrum after taking the square root of each component. It lacks most of the pitch information that is removed by the heavy cepstral smoothing during feature extraction. The technical details of the presented steps can be found in other publications [35][36][54]. The magnitude spectrum's counterpart, the phase spectrum, to start the inversion to the time domain via the Inverse Fast Fourier Transform (IFFT) is still missing. It was discarded during feature extraction and has to be estimated from the information present in the overlapping of frames. For this purpose, the iterative method introduced by Griffin and Lim [22] is used and the process is stopped when the average absolute difference (error) between two successive iterations of the magnitude spectrum is less than 4% of the average magnitude in the current spectrum (or 100 iterations are reached, whatever happens first).

This technical setting represents an extension of the approach taken by Demuynck et al. [13]. The novelty of our approach is that the modeling stage is also taken into account and that it is applied to the domain of speaker recognition, shifting the focus to speaker-related features.

2.4 Results

This subsection presents the results of the human speaker grouping study. Both quantitative and qualitative results will be discussed. We start with the quantitative outcomes showing how "well" the participants did the job.

Table 1 contains some statistical measures: mean and standard deviation of the time (in minutes) used to solve the task, the number of clusters created, the number of correctly drawn connections between utterances (considered transitively) and the number of connections drawn overall (without considering transitivity). Furthermore, the probability for the two segments of each of the 7 speakers to be joined (also considering transitivity) is presented in the remaining columns. These are stated for human annotations and "random"[1] clustering for all three data sets. A third line per data set shows the result of a one-sided t-test (H_0: human figure equals random figure; H_1: human figure is better than random) in terms of the minimal α-level possible to reject the null hypothesis (or "no" if it cannot be rejected). The t-value is computed using a pooled variance due to the small sample size of 20 on the side of the human annotations. The results can be summarized as follows:

a Human performance improves from run to run as in-

[1]It is important to know whether the human results deviate from pure human guessing. But what is a guessed result on a clustering task, where both the number of clusters as well as the affiliations to clusters must be guessed and both choices interdepend? We observe that a human will never choose cluster sizes and numbers totally at random, but will follow some intuition like "there will be more than one and less than the maximally possible number of clusters" and "there must be clusters having a 'reasonable' number of members". Therefore, we take the distributions of numbers and sizes as created by the participants for each data set and draw the guessed numbers and sizes of clusters at random from them. The members of the created empty clusters are then picked at random (i.e., uniformly distributed) from the set of still unassigned utterances. In this Monte Carlo way, we simulate 10000 independent random clustering runs per data set and present their outcome, getting results that are less purely random but more like human guessing.

Table 2: Performance of human and random clustering in terms of different figures of merit

means	dataset	rec_o	$prec_o$	MR	acp	asp	$purity$	γ	I_{BBN}	DER
human $\mu(\sigma)$	1	0.52 (0.11)	0.57 (0.16)	0.48 (0.11)	0.57 (0.16)	0.71 (0.12)	0.63 (0.09)	16.4 (8.37)	4.13 (1.78)	0.33 (0.1)
random $\mu(\sigma)$	1	0.41 (0.1)	0.44 (0.13)	0.59 (0.1)	0.51 (0.1)	0.57 (0.07)	0.53 (0.06)	18.3 (6.52)	3.42 (1.25)	0.35 (0.1)
human wins?	1	0.0005	0.0005	0.0005	0.005	0.0005	0.0005	0.1	0.01	no
human $\mu(\sigma)$	2	0.62 (0.19)	0.63 (0.19)	0.38 (0.19)	0.64 (0.18)	0.74 (0.14)	0.68 (0.15)	12.7 (7.18)	5.39 (2.5)	0.28 (0.14)
random $\mu(\sigma)$	2	0.42 (0.1)	0.44 (0.12)	0.58 (0.1)	0.52 (0.09)	0.56 (0.07)	0.54 (0.06)	16.4 (4.48)	3.63 (1.1)	0.34 (0.1)
human wins?	2	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.005
human $\mu(\sigma)$	3	0.96 (0.08)	0.98 (0.05)	0.04 (0.08)	0.98 (0.05)	0.96 (0.08)	0.97 (0.06)	0.9 (1.8)	10.1 (0.88)	0.02 (0.05)
random $\mu(\sigma)$	3	0.42 (0.1)	0.45 (0.11)	0.58 (0.1)	0.56 (0.06)	0.54 (0.05)	0.55 (0.05)	12.74 (1.67)	4.19 (0.76)	0.32 (0.1)
human wins?	3	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005

dicated by more correct connections using less time as well as cluster and connection numbers approaching the real values (7/7) with less variability.

b Nevertheless, individual speakers differ in how well their voices could be recognized – FELC0 and MREB0 have a consistently higher probability of being grouped correctly by humans among all data sets, whereas there is a consistently lower probability for FAKS0.

c Human results deviate positively (i.e., are better) from the random outcomes with a confidence of at least 99.5% in terms of the number of correct connections drawn and also in the probability of grouping together the correct utterances for almost all speakers on all three data sets.
Due to the fact that the random cluster sizes and numbers of clusters were drawn from the discrete distribution per data set created by the human participants, those figures do not deviate significantly; the small deviation is because the distributions are not Gaussian but somehow skewed and multimodal, so that with increased sample size in the random case (10000 as compared to 20) it becomes obvious that mean and standard deviation are inappropriate measures to describe these distributions.

Table 2 evaluates the achievements of human and random annotators in terms of several figures of merit, as defined in the survey by Kotti et al. [28] (except for overall recall and precision): overall recall rec_o and overall precision $prec_o$ are extensions of the usual recall and precision measures of the information retrieval community used for the task of clustering; they give the ratio of utterances being in the correct or a fitting cluster, respectively, where a fitting cluster is one where the utterances' speaker is in the majority, whereas the correct cluster is the biggest one of this speaker. The misclassification rate MR gives the likelihood of an utterance not getting assigned to the correct cluster. Average cluster purity acp is the likelihood of the utterances in one cluster really belonging together, whereas average speaker purity asp is the likelihood of utterances being assigned to speaker x really being spoken by x; the $purity$ is the geometric mean of both. The Rand index γ is an unnormalized number decreasing with the number of correctly clustered utterances, whereas the BBN metric I_{BBN} increases (unnormalized, too) with the number of big, pure clusters. The diarization error rate DER finally depicts the ratio of *samples* assigned to the wrong speaker, including speaker error time, missed speaker time, and false alarm speaker time (but due to the fact that we only evaluate clustering, the latter two sources of error are eliminated here).

There are several possibilities of selecting entities for computing figures of merit: audio samples would be the most accurate way, then segments (as created by silence detection, which would reduce to sentences here) or utterances (i.e., entire files consisting of concatenated sentences in our database). We have chosen utterances because they reflect most naturally what a human considers to be a good achievement; sample- or segment-level computation would introduce biases towards (or against) the longer segments, whereas this way each utterance is weighted equally. Several observations are noteworthy:

d Confirming the statistical results above, the human figures of merit get consistently and strictly monotonically better across runs.

e There are three important exceptions to the fact that all other human results are with at least 99.5% confidence better than random: for γ, I_{BBN} and DER on dataset 1, there is considerably less or no basis to deduce that they deviate from pure guessing; all three measures have in common (in contrary to the other ones) that they evaluate clustering in total.

f Average human performance on dataset 3 (the natural, "easy" one) is not perfect, but almost perfect.

g The biggest increase in performance seems to be between run (dataset) 2 and 3 (the latter is nearly perfect), which is on average 4.72 (with a standard deviation of 1.46) times greater than the gain between run 1 and 2 (the former is nearly guessing). But careful analysis reveals: the standard deviation for all measures in run 2 is considerably higher than for run 1 (and, less important, run 3). Looking inside the individual participant's results (not shown here for space reasons) shows that there are two groups of participants that are distinct: the major group (17 from 20 persons) gives results as indicated by Table 2; but two subjects score perfectly, another one has made only one wrong connection. These three participants have in common that they nearly exhausted the given time limit (median of 30 minutes), in contrast to everyone in the first group (median of 14 minutes).

h In run 1, the top 3 participants (there is no clear division of the subjects into groups) in terms of Rand index also use considerably more time (median of 30 minutes) than the rest (median of 20 minutes).

i There is no correlation of a participant's individual properties (sex, age or time taken to complete a task) with scoring considerably better or worse in any other run.

The qualitative results exhibit "how" the participants accomplished each task. They are assembled from the free text fields for each run on the assessment sheet. Due to the nature of free text, phrasing among the participants differs (and many have not commented on all of the indirect inquiries). Nevertheless, the results are very homogeneous, as confirmed by several oral inquiries consulting randomly selected participants. Table 3 reports on the features used by the participants on the different data sets. The popularity values display how often respective features are mentioned by the participants after summarizing similar references. Some broader categories include more detailed features besides and beyond the pure meaning of their names after summarization: rhythm/velocity includes concentrating on frequency *changes* as well as the accentuation and use of pauses; pitch includes separating "high" from "low" voices, which extends the psychoacoustical notion of pitch [38] to a broader view of main spectral components; timbre/sound includes articulation, accent, speaking style and intonation. The following findings are noticeable:

j With the data set's number, the usage of features that allow a vivid perception of the voice increases. It basically starts on dataset 2 with the mentioning of imagining the speaker behind the voice and the use of gender detection (although other participants state that this is impossible on this data set) and is used on dataset 3, where participants even clustered based on inferred attractive appearance of female speakers.

k An appeal to the normal human speech perception mode (i.e., holistic hearing) that is distinct in nature from perceiving other sounds being judged based on simple patterns and features as described by Moore [38] is only made for dataset 3.

l The features used for dataset 1 mostly confused the participants: rhythm/velocity as well as timbre/sound do not convey speaker-related information in dataset 1 because any inter-frame relationships are purely random by design.

m Regarding the methodology, the participants broadly adopt a systematic way of pairwise comparison of voices by adding them up to clusters until a certain threshold of dissimilarity is reached. The process then restarts with the next free utterance.

n In some cases, a multi-pass scheme that first skims a whole data set and then clusters utterances based on a process of elimination can be observed.

o For dataset 3, a hierarchical scheme that first presorts utterances by gender (a cue described as most helpful by several participants) before building groups can be observed.

p Some participants do not use any systematic strategy on dataset 3 but just "do it naturally".

The findings from the self-assessment of participants are summarized as follows:

q The quantitative results from above are largely confirmed – judgments are between "impossible" and "very unsure" on dataset 1 and do not vary much for dataset 2, where the range is from "very unsure" to "mediocre" with an emphasis on the first one. For dataset 3, the self-assessment is "quite correct" and predominantly "sure".

r The self-assessment for the second data set partly contradicts the measured clustering performance in that even the participants of the group of well-doing subjects do not regard themselves as being able of clustering the data.

Table 3: Popularity of human-used features

feature	#dataset 1	#dataset 2	#dataset 3
rhythm/velocity	7	11	8
pitch	7	11	7
timbre/sound	3	6	14
perceived gender	0	2	13
perceived age	0	0	5
visual imagination	0	1	3
volume	2	1	0
nasalization	0	1	0
holistic judgment	0	0	1

3. HARNESSING THE RESULTS

The aim of this paper is to identify speaker clustering stages that need to be improved and the order in which these improvements have to take place such that a maximum performance gain is obtained. The findings of Section 2 are now evaluated with respect to this aim.

3.1 Interpretation

First, our results of Section 2.4 confirm the choices made earlier in this paper as well as the popularity of common techniques:

a The results 2.4.i and the homogeneity of the qualitative results indicate that the choice of the set of participants is appropriate.

b The results 2.4.m to 2.4.p indicate that humans apply, in the absence of the subconscious speech mode used when everything is familiar, a way of accomplishing the task of grouping that resembles the algorithm in an automatic hierarchical clustering system: evaluating pairwise distances, grouping the closest clusters until a termination criterion is met, guided by any available additional information like sex. This justifies the omission of the clustering stage in the list of potential stages for improvement.

c Several results give evidence that the used MFCC features capture speaker-specific information quite well: 2.4.c and 2.4.e show that humans clearly perform better than guessing on dataset 2, and 2.4.g and 2.4.h suggest that achieving even better results on unfamiliar data might be a concentration issue rather than a matter of missing cues in the features. Moreover, Rose reports on experiments showing that human performance normally nearly doubles when exposed to familiar voices as opposed to unfamiliar ones [50, p. 103]. We argue that this performance loss in the presence of unfamiliarity is even more present when the sound itself is unusual.

d As indicated by 2.4.c, modeling is effective in the sense that GMMs even contain human-exploitable speaker-related information (although the main statement of 2.4.e needs further treatment below).

e The last two remarks allow us to conclude that humans are capable of analyzing this kind of sounds in principle, which supports our biomimetic approach. Further justification comes from Furui [20] who points out that breakthroughs will rather come from a better understanding of speech and the way it is produced and *perceived* rather than from mere improvements in statistical pattern recognition; and from Wu et al. [65] who also use the opportunity of learning from human speech processing abilities.

Second, there is evidence for a specific answer to our opening question. From 2.4.a, 2.4.d and 2.4.q it is clear that it is appropriate to view the pattern recognition chain as a process of information compression – exploitable as well as useless information with respect to speaker identity is abolished in each step. 2.4.g introduces our main argument by showing where the most useful information disappears: it is in the modeling stage. At first glance, 2.4.g seems to contradict this finding, but even though the figures of merit deviate more among dataset 2 and 3 than between dataset 1 and 2, there is a fundamental difference between both transitions. From dataset 3 to 2, average human performance drops from nearly perfect to below what is considered acceptable for a clustering system; but there is still this group of three candidates scoring nearly perfectly also on the audible features. On dataset 1, however, the complete clustering performance for all participants tends towards guessing (2.4.e) and no one considers himself able of accomplishing the task in contrast to dataset 2 (2.4.q). The fundamental difference is this: what is difficult on the audible features becomes impossible on the audible models. This does not contradict the conclusion that exploitable information is found in the models; individual voices can still be recognized quite well even on dataset 1 (2.4.b) – but the task of clustering dataset 1 *as a whole* becomes impossible.

What is it that produces this frontier between the feature extraction and modeling stage? 2.4.j suggests that participants find no features within audible models that help making the "voices" vivid. Table 3 shows what these features are: the timbre or sound of a voice, as well as the rhythm and velocity of the stream of speech (the latter ones have also been used by participants on dataset 1, but in a wrong way, see 2.4.l). These features have in common that they are essentially supra-frame based – they are not grasped in a single instant of time, but the sensation needs an evolution of frames to emerge. What is crucially missing in the modeling stage is an account for time.

Another point for optimization lies in the feature extraction stage: Participants found the preclassification of utterances by perceived gender most helpful (2.4.o), and gender is strongly correlated with the pitch of a voice. A sensation of pitch, though, is largely eliminated by design in MFCCs.

To summarize, we find that our features include what it takes to identify a voice (at least for a human analyst; no proposition is made that to be useful for machines, it might not be necessary to make certain parts of the vectors' content more explicit). But they would benefit from providing further cues for gender detection, i.e., pitch (or its acoustic correlate, F_0). But this improvement must be succeeded by an enhancement of the applied models to incorporate an account for the temporal succession of frames without modeling speech instead of a voice. This is the area with the highest potential for improvement.

3.2 Discussion

There are several promising approaches for finding better features, e.g. by Pachet and Roy [40], Thiruvaran et al. [59] or Prasanna et al. [41]. But until modeling is capable of capturing the fundamental relationships among individual vectors, these approaches will not yield what might be expected. This is also true for examples of accompanying MFCCs with pitch (or better: F_0) as done by Lu and Zhang [29], whose results are not better than those of comparable approaches [28]. Nevertheless, F_0 is an important feature also for forensic phoneticians, from whom striving for a better understanding of speech instead of improving technical solutions can most likely be expected – it is the most often mentioned single feature in Rose' book [50, pp. 41, 161/162, 246, 249/250]. However, apart from spectral (cepstral) features, all other features mentioned there have one thing in common: they exploit the temporal coherence of speech. Those features are: temporal factors (p. 113), breath patterns (p. 113), speaking tempo (p. 115), syllable grouping (p. 133), speech rate (p. 169) and hesitation (p.172).

Lindblom et al. use the temporal context of spectral frames to improve the extraction of formant center frequencies and conclude that the "temporal fine structure of the signal plays a very significant role [...] in speech perception" [30]. In a current attempt to identify future traits of research in biometrics, Schouten et al. put the demand for context inclusion on top of their list of 19 urgent topics [52]. The need for and the realization of the integration of temporal context has also recently been discovered by Aucouturier [2] and Joder et al. [25], respectively, for the field of music information retrieval. It follows that there is a widespread awareness of the importance of time-based information for audio processing.

The easiest way of modeling time dependencies is by accompanying feature vectors with their temporal derivatives of first and second order (δ and $\delta\delta$ features). Malegaonkar et al. show that this has some potential [32], but the positive effect is not consistently observable [27]. Another approach lies in the area of prosody modeling for speaker recognition: approaches there try to capture intonation, stress, rhythm and velocity of speech by modeling the trajectories of F_0 and/or short time energy over the duration of syllable-like units (50–100 ms according to Rose [50, p. 167]). Adami gives a good overview [1] and presents his approach of modeling the joint distribution of pitch- and energy-gestures along with their durations via bigrams. A gesture lasts until either the pitch- or energy-contour changes direction and is quantized into one of 5 states encoding the joint pattern of rise and descent of the two features. Mary and Yegnanarayana presegment the speech by detecting vowel onset points (VOP) before extracting mean-, peak- and change in F_0, peak-distance to VOP, amplitude- and duration-tilt and finally change in log-energy per segment as features for prosodic behavior and modeling them via auto-associative neural nets [33]. Further systems come, for example, from Reynolds et al. [44][43], Ferrer et al. [18] and Soenmez et al. [55]. They all have in common that the prosodic features

and models complement conventional (cepstrum-based) systems and improve the final result; that they are robust to noise and other variations; and that they need much data for training and testing in the region of several minutes.

Modeling prosodic speaker-dependent information heads into the right direction, but does not cover completely what is claimed by our study. First, not all of the features mentioned by the participants fall into the category of prosody: timbre and sound, for example, account for more than what is covered by energy- and pitch contours; they emerge with time, but likely with the time evolution of gross spectral shapes instead of just amplitude and fundamental frequency. Second, the features used by our participants could readily be evaluated with small amounts of training and test data (some participants reported to have used only the first 5–10 seconds to judge an utterance), whereas current prosodic systems suffer from the need for vast data consumption, as pointed out by Chen et al. [10]. Rose seems to bridge this gap with the following suggestion: the quality of a voice is best viewed in contrast to (or deviation from) an idealized neutral vocal apparatus configuration [50, p. 279] and the analysis might better focus on individual outstanding events rather than on global averages [50, p. 73]. A human listener with general knowledge of how speech sounds can find those outstanding speaker-specific sounds in a short utterance and reliably recognizes the voice based on them. Current prosodic systems do not possess this general knowledge and hence cannot find the few interesting parts of the signal, eventually needing more data for compensation.

4. IMPLEMENTATION AND RESULTS

Several ways are imaginable to implement the exploitation of time- and pitch information in the spirit of our results. In this section, we present an implementation of a speaker clustering system incorporating this kind of information.

Our "time model" replaces the GMM in our diarization framework presented in Section 1; everything else is left unchanged. The following new processing steps are incorporated in the time model:

- Speaking rate normalization
- Transformation of basic features to trajectories
- Estimation of the support of the trajectory's distribution in time and frequency
- Comparison of different trajectory models

The central idea is trajectory modeling: feature vectors of one utterance are not independent of each other, but belong to their temporal context. This context can be grasped by concatenating several subsequent single frames to a "context vector". It depends on the viewpoint whether this can be considered as improving the features instead of the modeling – in our implementation, the modeling stage receives a set of feature vectors in their original order that is then exploited further, hence we speak of improving the modeling stage. Previous approaches to trajectory modeling include the work of Chengalvarayan and Deng [11], Saul and Rahim [51], Vlachos et al. [63] or Chandra Sekhar et al. [53]. We deviate from their approaches in the way we create, model and/or compare trajectories.

We take the ordered sequence of 19-dimensional MFCC feature vectors representing a single utterance as described in Section 1, enriched with the F_0 contour extracted via the RAPT algorithm [57], as our basic features and input to our time model. Each dimension is normalized to the range [0..1] using the min/max values found on all the TIMIT data.

Then, the speaking rate is normalized so that the same sound uttered in different tempi results in the same sequence of feature vectors. We perform this by first clustering the frames into $\frac{2T}{3}$ clusters via k-means, where T is the number of feature vectors in the utterance under consideration (this way, speaking rate normalization works adaptive). The factor of 66% has been found optimal in informal listening experiments. Each vector is then replaced with its centroid, and a sequence of identical centroids in the feature set is cut to length one, thus reliably shortening stretched sounds.

Then, 13 subsequent vectors are concatenated to form one context vector. This corresponds to a syllable length of 130 ms and is found to best capture speaker specific sounds in informal listening experiments over a range of 32–496 ms (in intervals of 16 ms). Our context vector step is one original frame, i.e., 10 ms. This way, two subsequent trajectories share $\frac{23}{24}$ identical speech samples (one frame difference, and frames have 50% overlap), such that the time-/frequency-information is spread into different corners of the 260-dimensional context vector space. This makes it more probable for a differently aligned context vector in the test phase to be recognized. Experiments showed that the remaining 5 ms possible displacement leads to very similar context vectors on otherwise identical data.

The set of context vectors of one utterance is then fed into a one-class SVM [58] training step. Using only positive examples to identify the $100{\cdot}(1{-}\nu)\%$ densest data points, it can (in contrast to a GMM) handle very high dimensional data. We used the implementation available in LibSVM [9] in conjunction with the RBF kernel. For all the speaker models, a common outlier factor of $\nu = 0.4825$ has been found effective; for the γ parameter of the SVM, we adopt a grid search optimization framework for each training set/model separately, using 5-fold cross validation in 25 logarithmically spaced steps between the minimum and maximum pairwise distances of all trajectories in the set. This individual parameter search is mainly responsible for the increased runtime, but appears to be crucial for the result.

After having built a time model for each utterance, the clustering procedure is applied using the Cross Likelihood Ratio (CLR) [28] as the metric between two models. CLR works considerably better in pretests than the Contrast Measure d_c presented by Desobry et al. [15][14], a direct measure between model parameters, and better than GLR as well. The likelihood of a set of MFCC+F_0 to a time model is computed as follows: feature vectors are transformed to context vectors using the methodology described above, and fed into the one-class SVM model. The ratio of positively classified trajectories is the desired likelihood.

This is a novel approach to voice modeling for the purpose of recognition. The processing steps inside the model as well as the various parameter settings originate from sound considerations but only preliminary experiments, leaving room for improvement. We have applied our time model to the clustering task on the reduced TIMIT data set with 40 speakers and 80 utterances that was used as an example when the baseline GMM approach starts to fail. Comparisons are made with the baseline MFCC/GMM approach presented in Section 1 and with several common approaches for time and

Table 4: Results

approach	runtime [m]	rec_o	$prec_o$	MR	DER
baseline	2.70	0.875	0.9875	0.125	0.04527
baseline+δ	4.95	0.35	0.35	0.65	0.5833
baseline+δ+$\delta\delta$	7.98	0.5	0.9875	0.5	0.1731
baseline+F_0	2.15	0.7375	0.9	0.2625	0.1551
baseline+δ+F_0	4.98	0.5125	0.5125	0.4875	0.4084
baseline+δ+$\delta\delta$+F_0	7.97	0.2875	0.2875	0.7125	0.6176
time model	523.13	0.9375	0.975	0.0625	0.01962

pitch exploitation, namely enhancing the MFCC vectors by δ, $\delta\delta$ and F_0 columns. All experiments have been carried out on a computer with 2 GB RAM and a Core2Duo processor at 2.4 GHz running our C++ based implementation under Fedora 10 Linux. Results are presented in Table 4.

First, the standard baseline system itself scores better than the enhanced baseline systems, which is in line with our previous reasoning, the results presented by Kotti et al. [28], and partly due to the curse of dimensionality letting GMMs perform poorly on higher-dimensional inputs [17]. Overall, our time model approach yields 56.66% and 50.00% relative DER and misclassification rate improvement over the standard baseline, respectively. These results indicate that time coherence exploitation (combined with pitch) as suggested by our study improves the performance of current speaker clustering systems.

5. CONCLUSIONS

The work presented in this paper is based on the observation that speaker clustering (diarization) approaches work considerably less satisfactory than approaches for the related tasks of speaker verification and identification. Therefore, we have presented a study to answer the following two questions by means of observing human behavior in a speaker clustering task: (a) where in the processing chain of speaker clustering has an improvement to take place to maximally improve the final outcome? (b) How does this improvement look like qualitatively?

The interpretation of our results has shown that it is the stage of modeling that bears the highest potential: the inclusion of temporal context information among feature vectors is what is crucially missing there. Furthermore, the inclusion of pitch information into feature vectors (in order to enable systems to better exploit gender information) is found to be a subordinate improvement – it will only have an effect when the major problem within modeling has been solved.

These results have lead to an implementation of a speaker clustering system that demonstrates the validity of our approach by outperforming common MFCC/GMM-based approaches on the reduced TIMIT benchmark with a relative improvement of 56.66% DER and 50.00% misclassification rate, respectively.

Two things should be noted about our approach: on the one hand, its design allows improvements in speaker clustering systems – time coherence e.g. clearly is a currently unexploited source of important information, and MFCCs modeled by GMMs will certainly not score above some glass ceiling in the spirit of Aucouturier and Pachet [3]. On the other hand, the biomimetic approach is not the only possible way to determine areas of improvement - other approaches may certainly be discovered.

There are several questions for future work: is the time succession of frames best grasped by concatenating several frames together? What are good conditions and parameter settings for the one-class SVM model and how can they be found? How can, according to Rose [50, p. 73], the outstanding trajectories of a speaker be found and technically exploited? How can the increased runtime of the time model approach be improved? Finally, how can the entire temporal context be considered, just as in the popular forensic phonetic method of analyzing spectrograms in a Gestalt-based manner [50, p. 116]?

Acknowledgments

This work is funded by the Deutsche Forschungsgemeinschaft (German Research Foundation, SFB/FK615, Project MT). We would like to thank our probands participating in the study, Martin Schwalb for fruitful discussions and Zhengyou Zhang for his helpful comments.

6. REFERENCES

[1] A. G. Adami. Modeling Prosodic Differences for Speaker Recognition. *Speech Communication*, 49:277–291, 2007.

[2] J.-J. Aucouturier. A Day in the Life of a Gaussian Mixture Model: Informing Music Pattern Recognition with Psychological Experiments. *Journal of New Music Research*, submitted, 2009.

[3] J.-J. Aucouturier and F. Pachet. Improving Timbre Similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

[4] Y. Bar-Cohen. *Biomimetics: Biologically Inspired Technologies*. CRC Press, Boca Raton, FL, USA, 2006.

[5] H. Beigi, S. Maes, and J. Sorensen. A Distance Measure Between Collections of Distributions and its Application to Speaker Recognition. In *IEEE Proc. of ICASSP*, volume 2, pages 753–756, 1998.

[6] J. Benesty, M. M. Sondhi, and Y. Huang. *Springer Handbook of Speech Processing*. Springer, Germany, 2008.

[7] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, USA, 2006.

[8] J. P. Campbell. Speaker Recognition: A Tutorial. *Proceedings of the IEEE*, 85:1437–1462, 1997.

[9] C.-C. Chang and C.-J. Lin. *LIBSVM: A Library for Support Vector Machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[10] Z.-H. Chen, Y.-F. Liao, and Y.-T. Juang. Prosody Modeling and Eigen-Prosody Analysis for Robust Speaker Recognition. In *Proc. IEEE Int. Conf. Acoust. Speech & Signal Proc. ICASSP'05*, pages I–185–I–188, 2005.

[11] R. Chengalvarayan and L. Deng. Speech Trajectory Discrimination Using the Minimum Classification Error Learning. *IEEE Transactions on Speech and Audio Processing*, 6(6), 1998.

[12] S. Davis and P. Mermelstein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28:357–366, 1980.

[13] K. Demuynck, O. Garcia, and D. V. Compernolle. Synthesizing Speech from Speech Recognition Parameters. In *Proc. International Conference on Spoken Language Processing, Jeju Island, Korea*, volume II, pages 945–948, 2004.

[14] F. Desobry, M. Davy, and C. Doncarli. An Online Kernel Change Detection Algorithm. *IEEE Transactions on Signal Processing*, 53(8), 2005.

[15] F. Desobry, M. Davy, and W. J. Fitzgerald. A Class of Kernels for Sets of Vectors. In *Proceedings of ESANN'2005*, pages 461–466. MIT Press, 2005.

[16] M. Faundez-Zanuy and E. Monte-Moreno. State-of-the-Art in Speaker Recognition. *IEEE Aerospace and Electronic Systems Magazine*, 20:7–12, 2005.

[17] B. Fergani, M. Davy, and A. Houacine. Speaker Diarization using One-Class Support Vector Machines. *Speech Communication*, 50:355–365, 2008.

[18] L. Ferrer, H. Bratt, V. R. R. Gadde, S. Kajarekar, E. Shriberg, K. Sönmez, A. Stolcke, and A. Venkataraman. Modeling Duration Patterns for Speaker Recognition. In *Proceedings of EUROSPEECH*, pages 2017–2020, 2003.

[19] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall. The DARPA Speech Recognition Research Database: Specification and Status. In *Proceedings of the DARPA Speech Recognition Workshop, Report No. SAIC-86/1546, February 1986, Palo-Alto*, 1986.

[20] S. Furui. 50 Years of Progress in Speech and Speaker Recognition. In *Proc. SPECOM 2005, Patras, Greece*, pages 1–9, 2005.

[21] B. Goertzel and C. Pennachin. *Artificial General Intelligence*. Springer, Berlin, Heidelberg, Germany, 2007.

[22] D. W. Griffin and J. S. Lim. Signal Estimation from Modified Short-Time Fourier Transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32:236–243, 1984.

[23] K. J. Han, S. Kim, and S. S. Narayanan. Strategies to Improve the Robustness of Agglomerative Hierarchical Clustering Under Data Source Variation for Speaker Diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 16:1590–1601, 2008.

[24] H. Jin, F. Kubala, and R. Schwartz. Automatic Speaker Clustering. In *Proc. of the DARPA Speech Recognition Workshop*, pages 108–111, 1997.

[25] C. Joder, S. Essid, and G. Richard. Temporal Integration for Audio Classification With Application to Musical Instrument Classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 17:174–186, 2009.

[26] D. E. Knuth. *The Art of Computer Programming, Volume 2: Seminumerical Algorithms, 3rd Edn.* Addison Wesley, 1998.

[27] M. Kotti, E. Benetos, and C. Kotropoulos. Computationally Efficient and Robust BIC-Based Speaker Segmentation. *IEEE Transactions on Audio, Speech, and Language Processing*, 16:920–933, 2008.

[28] M. Kotti, V. Moschou, and C. Kotropoulos. Speaker Segmentation and Clustering. *Signal Processing*, 88:1091–1124, 2008.

[29] H.-J. Z. Lie Lu. Unsupervised Speaker Segmentation and Tracking in Real-Time Audio Content Analysis. *Multimedia Systems*, 10:332–343, 2005.

[30] B. Lindblom, R. Diehl, and C. Creeger. Do 'Dominant Frequencies' Explain the Listener's Response to Formant and Spectrum Shape Variations? *Speech Communication*, 2008.

[31] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava. Speech and Language Technologies for Audio Indexing and Retrieval. *Proceedings of the IEEE*, 88:1338–1353, 2000.

[32] A. Malegaonkar, A. Ariyaeeinia, P. Sivakumaran, and S. Pillay. Discrimination Effectiveness of Speech Cepstral Features. *Lecture Notes in Computer Science*, 5372:91–99, 2008.

[33] L. Mary and B. Yegnanarayana. Extraction and Representation of Prosodic Features. *Speech Communication*, 2008.

[34] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier. Step-by-Step and Integrated Approaches in Broadcast News Speaker Diarization. *Computer Speech and Language*, 20:303–330, 2006.

[35] B. Milner and X. Shao. Speech Reconstruction from Mel-Frequency Cepstral Coefficients using a Source-Filter Model. In *International Conference on Spoken Language Processing (ICSLP)*, pages 2421–2424, 2002.

[36] B. Milner and X. Shao. Clean Speech Reconstruction from MFCC Vectors and Fundamental Frequency using an Integrated Front-End. *Speech Communication*, 48:697–715, 2006.

[37] T. M. Mitchell. *Machine Learning*. WCB/McGraw-Hill, 1997.

[38] B. C. J. Moore. *Psychology of Hearing, Fifth Edition*. Elsevier Academic Press, London, UK, 2004.

[39] A. Morris, D. Wu, and J. Koreman. GMM based Clustering and Speaker Separability in the TIMIT Speech Database. Technical Report Saar-IP-08-08-2004, Saarland University, 2004.

[40] F. Pachet and P. Roy. Exploring Billions of Audio Features. In Eurasip, editor, *Proceedings of CBMI 07*, pages 227–235, 2007.

[41] S. M. Prasanna, C. S. Gupta, and B. Yegnanarayana. Extraction of Speaker-Specific Excitation Information from Linear Prediction Residual of Speech. *Speech Communication*, 48:1243–1261, 2006.

[42] M. Przybocki and A. Martin. NIST Speaker Recognition Evaluation Chronicles. In *Proceedings in Odyssey 2004*, 2004.

[43] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang. The SuperSID Project: Exploiting High-Level Information for High-Accuracy Speaker Recognition. In *Proc. IEEE Int. Conf. Acoust. Speech & Signal Proc. ICASSP'03*, pages IV–784–IV–787, 2003.

[44] D. Reynolds, W. Campbell, T. Gleason, C. Quillen, D. Sturim, P. Torres-Carrasquillo, and A. Adami. The 2004 MIT Lincoln Laboratory Speaker Recognition System. In *Proc. IEEE Int. Conf. Acoust. Speech & Signal Proc. ICASSP'05*, pages I–177–I–180, 2005.

[45] D. Reynolds and P. Torres-Carrasquillo. The MIT Lincoln Laboratory RT-04F Diarization Systems: Applications to Broadcast News and Telephone Conversations. In *NIST Rich Transcription Workshop November 2004*, 2004.

[46] D. Reynolds and P. Torres-Carrasquillo. Approaches and Applications of Audio Diarization. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing 2005*, volume 5, pages V-953–V-956, 2005.

[47] D. A. Reynolds. Speaker Identification and Verification using Gaussian Mixture Speaker Models. *Speech Communication*, 17:91–108, 1995.

[48] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10:19–41, 2000.

[49] D. A. Reynolds and R. C. Rose. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. *IEEE Transactions on Speech and Audio Processing*, 3:72–83, 1995.

[50] P. Rose. *Forensic Speaker Identification*. Taylor & Francis, London and New York, 2002.

[51] L. Saul and M. Rahim. Markov Processes on Curves for Automatic Speech Recognition. In *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*, pages 751–757. MIT Press, 1999.

[52] B. Schouten, M. Tistarelli, C. Garcia-Mateo, F. Deravi, and M. Meints. Nineteen Urgent Research Topics in Biometrics and Identity Management. *Lecture Notes in Computer Science*, 5372:228–235, 2008.

[53] C. C. Sekhar and M. Panaliswami. Classification of Multidimensional Trajectories for Acoustic Modeling Using Support Vector Machines. In *Proceedings of ICISIP'04*, pages 153–158, 2004.

[54] S. W. Smith. *Digital Signal Processing - A Practical Guide for Engineers and Scientists*. Newnes, USA, 2003.

[55] M. K. Soenmez, L. Heck, M. Weintraub, and E. Shriberg. A Lognormal Tied Mixture Model of Pitch for Prosody-Based Speaker Recognition. In *Proceedings of EUROSPEECH*, pages 1391–1394, 1997.

[56] T. Su and J. G. Dy. In Search of Deterministic Methods for Initializing K-Means and Gaussian Mixture Clustering. *Intelligent Data Analysis*, 11:319–338, 2007.

[57] D. Talkin. A Robust Algorithm for Pitch Tracking (RAPT). In W. B. Klejin and K. K. Paliwal, editors, *Speech Coding and Synthesis*, chapter 3, pages 495–518. Elsevier Science, Amsterdam, NL, 1995.

[58] D. M. J. Tax. *One-Class Classification - Concept-Learning in the Absence of Counter-Examples*. PhD thesis, Technische Universteit Delft, 2001.

[59] T. Thiruvaran, E. Ambikairajah, and J. Epps. Group Delay Features for Speaker Recognition. In *6th International Conference on Information, Communications & Signal Processing*, pages 1–5, 2007.

[60] S. E. Tranter and D. A. Reynolds. An Overview of Automatic Speaker Diarization Systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:1557–1565, 2006.

[61] W.-H. Tsai, S.-S. Chen, and H.-M. Wang. Automatic Speaker Clustering using a Voice Characteristic Reference Space and Maximum Purity Estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15:1461–1474, 2007.

[62] D. A. van Leeuwen, A. F. Martin, M. A. Przybocki, and J. S. Bouten. NIST and NFI-TNO Evaluations of Automatic Speaker Recognition. *Computer Speech and Language*, 20:128–158, 2006.

[63] M. Vlachos, G. Kollios, and D. Gunopulos. Discovering Similar Multidimensional Trajectories. In *Proceedings of ICDE'02*, pages 673–684, 2002.

[64] D. Wu. *Discriminative Preprocessing of Speech: Towards Improving Biometric Authentication*. PhD thesis, Saarland University, 2006.

[65] D. Wu, J. Li, and H. Wu. α-Gaussian Mixture Modelling for Speaker Recognition. *Pattern Recognition Letters*, 2009.

[66] S. Zhang, W. Hu, T. Wang, J. Liu, and Y. Zhang. Speaker Clustering Aided by Visual Dialogue Analysis. In *PCM 2008, Lecture Notes on Computer Science*, volume 5353, pages 693–702, 2008.