Multimedia Tools and Applications manuscript No.
(will be inserted by the editor)

Lexical Speaker Identification in TV Shows

Anindya Roy · Hervé Bredin · William Hartmann · Viet Bac Le · Claude Barras · Jean-Luc Gauvain

Received: date / Accepted: date

Abstract It is possible to use lexical information extracted from speech transcripts for speaker identification (SID), either on its own or to improve the performance of standard cepstral-based SID systems upon fusion. This was established before typically using *isolated* speech from *single* speakers (NIST SRE corpora, parliamentary speeches). On the contrary, this work applies lexical approaches for SID on a different type of data. It uses the REPERE corpus consisting of *unsegmented multiparty* conversations, mostly debates, discussions and Q&A sessions from TV shows. It is hypothesized that people give out clues to their identity when speaking in such settings which this work aims to exploit. The impact on SID performance of the diarization front-end required to pre-process the unsegmented data is also measured.

Four lexical SID approaches are studied in this work, including TFIDF, BM25 and LDA-based topic modeling. Results are analysed in terms of TV shows and speaker roles. Lexical approaches achieve low error rates for certain speaker roles such as anchors and journalists, sometimes lower than a standard cepstral-based Gaussian Supervector - Support Vector Machine (GSV-SVM) system. Also, in certain cases, the lexical system shows modest improvement over the cepstral-based system performance using score-level sum fusion.

To highlight the potential of using lexical information not just to improve upon cepstral-based SID systems but as an independent approach in its own right, initial studies on *crossmedia* SID is briefly reported. Instead of using

Anindya Roy, Hervé Bredin, William Hartmann, Claude Barras, Jean-Luc Gauvain,
LIMSI-CNRS
Tel.: +33 1 69 85 80 80
Fax : +33 1 69 85 80 88
E-mail: {roy, bredin, hartmann, barras, gauvain}@limsi.fr

Viet Bac Le,
Vocapia Research
Tel.: +33 1 60 14 97 73
Fax : +33 1 60 19 54 94
E-mail: levb@vocapia.com

speech data as all cepstral systems require, this approach uses Wikipedia texts to train lexical speaker models which are then tested on speech transcripts to identify speakers.

Keywords Lexical speaker identification · broadcast conversations · TFIDF · BM25 · speaker roles · classifier fusion · crossmedia learning · Wikipedia

1 Introduction

Traditional speaker identification (SID) systems use Gaussian mixture models (GMM) to approximate the distribution of cepstral features extracted from short speech frames of length 20 to 30 ms [20]. State-of-the-art systems build on top of this basic framework by concatenating the GMM mean vectors to form speaker-specific *supervectors*, typically classified using Support Vector Machines (SVM) [7] or processed using factor analysis [10] or I-vector analysis [9,26,1].

In contrast to these classical short-term cepstral-based approaches, development of automatic speech recognition (ASR) systems has led to approaches which model longer-term features dependent on the phonetic or lexical content of the automatic speech recognition output [12,6,2,19,18,27,30,4]. These include phonetic, prosodic, lexical or hybrid approaches. Among these approaches, the focus of this work is on *purely* lexical approaches which ignore all acoustic information after the speech has been transcribed to words. There are three main motivations behind this choice, as follows.

First, previous work on NIST SRE and parliamentary speech corpora has demonstrated the potential of such purely lexical approaches [27,4]. Second, such approaches can take advantage of easily available speech transcripts in the form of minutes of parliamentary speeches [4] or fan-made/commercially produced subtitles of movies and TV shows[1] to create lexical speaker models. Third, working in the lexical domain opens up another interesting possibility: person-related text information could be extracted from freely available Internet documents such as Wikipedia articles and news websites to train *lexical person models*. These models may later be used to identify speakers using ASR-derived speech transcripts. Note that the last setting (using Internet-derived text for training and ASR transcripts for testing) is noteworthy in that it does not need any prior speech from the speaker at all in the training phase (unlike *all* cepstral-based systems).

The acoustic modality of the REPERE multimodal corpus has been used for this work. This corpus consists of 7 different French TV shows. The interest in choosing this corpus is as follows. First, speakers in TV shows often give clues to their identity in the content of their speech. For example, they may name the show or the channel (for anchors and journalists). They may talk about the subject of their expertise or interest (for example, the upcoming

[1] Example: `www.opensubtitles.org`.

elections they will participate in, or the last movie they acted in). It is hypothesized that lexical approaches will be able to learn such speaker-discriminative information from speech content. Second, this corpus is comprised of unsegmented audio streams from TV shows with music and non-speech segments interspersed with speech segments from multiple speakers conversing together, unlike isolated speech from single speakers typically used in prior lexical SID studies. This gives an opportunity to study the behavior of lexical approaches in conjunction with an automatic diarization module to pre-process the data, a task which has not been performed before.

The first contribution of this paper is to propose an Information Retrieval (IR)-based approach, namely BM25 (OKAPI) for lexical SID and study its performance along with 3 existing representative lexical approaches, namely TFIDF, Markovian Term Weighting (MTW) and Latent Dirichlet Allocation (LDA). BM25 is well-established in the IR community but to the authors' knowledge, this is the first time it is applied to lexical SID. The second contribution is to carry out a set of contrastive experiments on the REPERE database using automatically *vs.* manually transcribed and diarized speech data for training and testing, and analyze the degradation of SID performance between the two cases. Third, a simple score-level sum fusion with a state-of-the-art acoustic system is proposed. Fourth, an initial study on crossmedia lexical SID is reported, which brings out the potential of such purely lexical approaches.

Note that a parallel approach for lexical SID in TV shows is to use lexical context around spoken names to classify the names between speaker, addressee and object [8, 29, 24]. On the contrary, this work does not depend on spoken names (hence neither on a Named Entity Recognizer), but rather analyzes the general lexical content of speech. However, it has the potential to be eventually combined with the former method.

The rest of the paper is organized as follows. Section 2 describes the pre-processing steps applied on the input audio stream. Section 3 describes our lexical speaker modeling framework. Section 4 describes the experiments and discusses the results. Section 5 briefly reports on a crossmedia lexical SID experiment. Section 6 concludes the paper.

2 Audio and text pre-processing

The input to the system is assumed to be a continuous audio stream recorded over the entire duration of a TV show. This is processed by the following steps in order.

2.1 Speaker diarization

The audio stream is first partitioned into speech and non-speech segments via GMM-based Viterbi segmentation [14]. The resulting speech segments are

then clustered into homogeneous speaker clusters via two steps: agglomerative clustering based on the BIC criterion to yield pure clusters followed by a second clustering step using cross-likelihood ratio (CLR) as the distance between clusters [3]. Since the REPERE corpus contains several episodes of TV shows and the same identifier should be associated to a given speaker across all the episodes, a first, local clustering step was followed by CLR clustering across all episodes [28]. The Diarization Error Rate (DER) for the system was around 16% on the REPERE development and test corpora.

2.2 Automatic speech recognition

A state of the art French speech-to-text transcription system [21] was then used to transcribe the audio in each speaker cluster. Decoding was carried out in a single real-time pass which produced a word lattice using cross-word, word-position dependent acoustic models, followed by consensus decoding with a 4-gram language model and pronunciation probabilities (35-phone set, 65K word vocabulary). At the end of this step, each speaker cluster was transcribed as a set of words. The case-insensitive word error rate for the system was 15.2% on the REPERE test corpus.

2.3 Text processing and tokenization

The set of words transcribed from each speaker cluster was processed to retain words with only alphabetic characters. Converting to lower case was found to improve SID performance. Initially, it was assumed that filler words such as "euh" and "hum" could be deleted and words repeated multiple times one after another could be replaced by a single instance of the word. However, contrastive experiments showed that such information slightly improved speaker ID performance. Hence, they were retained. It is common to use stemming but initial studies using TreeTagger[2] showed that it did not perform well on the noisy ASR output. Hence, it was not used further in this work.

Let the processed set of words transcribed from speaker cluster i be termed as document C_i. Words from all documents extracted from the REPERE training corpus were aggregated to form a vocabulary $\mathbf{v} = \{v_1, v_2, \cdots, v_{|\mathbf{v}|}\}$ of size $|\mathbf{v}| \approx 15\text{K}$, where each v_k represents a word. Each document C_i was then mapped to a word count vector $\mathbf{n}_i = [n_{i,1}, n_{i,2}, \cdots, n_{i,|\mathbf{v}|}]^{\mathrm{T}}$ where each $n_{i,k}$ contains the count of word v_k in C_i.

3 Lexical speaker modeling

Let the set of documents extracted from the REPERE training corpus be denoted by $\mathbf{C}_{\mathrm{tr}}$. Each of these documents represents a unique speaker associated

[2] www.cis.uni-muenchen.de/~schmid/tools/TreeTagger

with an identifier of the form `Firstname_LASTNAME` provided with the corpus. In this work, we used the word count vectors $\{\mathbf{n}_i\}$ computed from these documents as target speaker models. Given a count vector $\mathbf{n}_j$ computed from a previously unseen test document, a similarity score between $\mathbf{n}_i$ and $\mathbf{n}_j$ is computed in 4 different ways as follows.

3.1 Simple TF-IDF weighting

The Term Frequency-Inverse Document Frequency (TFIDF) weighting was used before for lexical SID, notably in [4]. There are multiple variations of this scheme in the literature [23]. The one which performed best in the SID task is described here. The IDF weight w_k for a word v_k in vocabulary $\mathbf{v}$ was defined as:

$$w_k = \log \frac{|\mathbf{C}_{\text{tr}}|}{|\mathrm{C}_i : \mathrm{C}_i \in \mathbf{C}_{\text{tr}} \wedge v_k \in \mathrm{C}_i|} \tag{1}$$

where $|\mathbf{C}_{\text{tr}}|$ denotes the number of documents in $\mathbf{C}_{\text{tr}}$ and $|\mathrm{C}_i : \mathrm{C}_i \in \mathbf{C}_{\text{tr}} \wedge v_k \in \mathrm{C}_i|$ denotes the number of documents in $\mathbf{C}_{\text{tr}}$ containing word v_k. Next, the TDIDF similarity score between a target model $\mathbf{n}_i$ and a test vector $\mathbf{n}_j$ was computed as the cosine distance between $\mathbf{n}_i$ and $(\mathbf{w} \diamond \mathbf{n}_j)$:

$$\mathrm{S}^{\text{TFIDF}}(i,j) = \frac{\mathbf{n}_i^T \cdot (\mathbf{w} \diamond \mathbf{n}_j)}{\|\mathbf{n}_i\| \, \|\mathbf{w} \diamond \mathbf{n}_j\|} \tag{2}$$

where $\mathbf{w} = [w_1, w_2, \cdots, w_{|\mathbf{v}|}]^{\mathrm{T}}$, $\diamond$ denotes element-wise multiplication and $\|\cdot\|$ denotes the Euclidean norm.

3.2 BM25 (OKAPI) weighting

Well-established in the IR community [23,17], the BM25 scheme has not been used before for lexical SID. We disinguish BM25 from simple TFIDF by the nonlinear mapping of raw word counts $\{n_{i,k}\}$ extracted from document C_i as follows:

$$n_{i,k}^{+} = \frac{(a+1) \cdot n_{i,k}}{a \cdot (1 - b + b \cdot \frac{l_i}{l}) + n_{i,k}} \tag{3}$$

where $n_{i,k}^{+}$ is the mapped count, a and b are parameters tuned to reduce the Identification Error Rate (ref. Section 4.1) on the development corpus, l_i is the number of words in document C_i and l is the mean document length of the corpus. Given two vectors $\mathbf{n}_i^{+}, \mathbf{n}_j^{+}$ consisting of mapped counts, the BM25 similarity score S^{BM25} is computed by Eq. 2 replacing $\mathbf{n}_i$ and $\mathbf{n}_j$ by $\mathbf{n}_i^{+}$ and $\mathbf{n}_j^{+}$ respectively.

3.3 Markovian term weighting (MTW)

This approach has yielded results comparable to BM25 in TREC evaluations [15] and has been applied previously to lexical SID [2] (interpreted there as MAP adapted models). The MTW similarity score between a target model $\mathbf{n}_i$ and a test vector $\mathbf{n}_j$ is computed as:

$$\mathrm{S}^{\mathrm{MTW}}(i,j) = \sum_{k=1}^{|\mathbf{V}|} n_{j,k} \cdot \log(\alpha \cdot p_{i,k} + (1-\alpha) \cdot p_k) \tag{4}$$

where $p_{i,k} = \frac{n_{i,k}}{\sum_{k'} n_{i,k'}}$ is the probability estimate of word v_k in document C_i, $p_k = \frac{\sum_{i'} n_{i',k}}{\sum_{i',k'} n_{i',k'}}$ is the probability estimate of word v_k over the entire training corpus and α is a parameter tuned on the development corpus.

3.4 Topic modeling (LDA)

Topic models have been used before for lexical SID as a viable altenative to direct word-based approaches [19,4]. We closely followed [4], using the Mallet implementation [25] of Latent Dirichlet Allocation (LDA) [5] to model each document as a distribution of topics, learnt on the REPERE training corpus. The optimal number of topics $\mathrm{N_T}$ was 20 for this work. Given the word count vector $\mathbf{n}_i$, the corresponding topic distribution $\mathbf{t}_i$ can be calculated. As in [4], symmetric KL-Divergence was used to calculate the LDA similarity score $\mathrm{S}^{\mathrm{LDA}}$ between a target model $\mathbf{n}_i$ and a test vector $\mathbf{n}_j$ as follows:

$$\mathrm{S}^{\mathrm{LDA}}(i,j) \equiv \mathrm{S}^{\mathrm{LDA}}(\mathbf{t}_i, \mathbf{t}_j) = \mathrm{D}(\mathbf{t}_i, \mathbf{t}_j) + \mathrm{D}(\mathbf{t}_j, \mathbf{t}_i) \tag{5}$$

where $\mathrm{D}(\mathbf{p}, \mathbf{q})$ is the standard KL-Divergence defined as:

$$\mathrm{D}(\mathbf{p}, \mathbf{q}) = \sum_{r=1}^{\mathrm{N_T}} p_r \log \frac{p_r}{q_r}. \tag{6}$$

Other lexical speaker modeling approaches The language model likelihood ratio framework [12] did not perform on this task, hence it is not reported here, although note that MTW may be considered a generalization of this framework and it did perform well (as shown in Section 4.3. Support Vector Machines (SVM) trained on TFIDF or n-gram count vectors [18,4] did not perform well possibly due to lack of sufficient training data, hence is not reported here too.

3.5 Decision-making

Given the set of similarity scores $\{\mathrm{S}(i,j)\}$ between count vector $\mathbf{n}_j$ extracted from test document C_j corresponding to a speaker cluster j and target models

$\{\mathbf{n}_i\}$ derived from the REPERE training corpus, the speaker cluster j is identified with the speaker hypothesis i_j^{HYP} which maximizes the similarity score $\mathrm{S}(i,j)$:

$$i_j^{\mathrm{HYP}} = \arg \max_{i:\mathrm{C}_i \in \mathbf{C}_{\mathrm{tr}}} \mathrm{S}(i,j) \tag{7}$$

The hypothesis i_j^{HYP} is assigned to all time instants t in the show within the set T_j of time segments corresponding to speaker cluster j.

$$i^{\mathrm{HYP}}(t) = i_j^{\mathrm{HYP}} \quad \forall t \in T_j \tag{8}$$

Note that scores are directly used in decision-making in this work. Score normalization for lexical systems will be studied in future.

4 Experimental evaluation

4.1 Database and protocol

Speaker identification experiments were carried out using the REPERE corpus under the framework of the REPERE Challenge[3] [13,16]. The corpus was collected by ELDA[4]. It is now available to all external consortia participating in the REPERE Challenge and will be publicly available in future. It consists of several episodes of 7 TV shows from 2 French channels broadcast through 2011-2012. The 7 shows are varied in nature, consisting of debates, discussions, news reports, and question time in the French parliament. Table 1 briefly describes each TV show and lists the tags (S1-S7) used to denote them henceforth in this work.

Inside the corpus, the audio streams are unsegmented, leading to the possibility of studying how automatic diarization affects performance. The corpus is partitioned into training (42 hrs), development (9 hrs) and test (9 hrs) corpora. Part of it is manually annotated in terms of speaker identity and transcribed text (training: 26 hours, development: 3 hrs, test: 3 hrs).

The number of manually identified speakers in the training corpus is 474. However, it was decided to retain only speakers who spoke for at least 10 seconds. This is because smaller training durations are too small to adapt GMMs in the cepstral-based GSV-SVM SID system that is used to compare with lexical SID systems (ref. Section 4.2). This resulted in 359 target speaker models. The number of manually identified reference speakers in the test corpus is 132. The number of speakers in common between target speakers in training corpus and reference speakers in test corpus is 63.

Table 2 lists the 5 speaker roles R1-R5 present in the corpus[5] and the average speaking duration of speakers with these roles (aggregated over all turns and episodes). Note that in train, only R1 and R2 speak for more than

[3] `www.defi-repere.fr` (in French)

[4] `www.elda.org`

[5] Role annotations were provided by ELDA with the corpus.

	TV Show	Brief description
S1	BFM Story	Daily interviews and debates on current affaires
S2	BFMTV Culture et Vous	Daily reports on cinema, music, art and literature
S3	LCP Ca Vous Regarde	Daily news show on politics and French parliament
S4	LCP Entre Les Lignes	Debates on current affaires between news editors
S5	LCP Info 13h30	Afternoon news and discussions on politics
S6	LCP Pile et Face	Debates between political personalities
S7	LCP Top Questions	Question time in the French Parliament

Table 1 Brief description of the 7 TV Shows in the REPERE database.

	Average duration of speaker cluster		
Speaker role	training	development	test
R1: Anchor	21 mins	4 min	4 min
R2: Journalist	14 mins	3 min	3 min
R3: Reporter	3 mins	2 min	1 min
R4: Guest	4 mins	2 min	2 min
R5: Others	1 min	1 min	1 min

Table 2 Role-wise breakup of average speaker cluster durations in REPERE training, development and test corpora, aggregated over all episodes. Note short durations for development and test.

5 minutes on average. Other roles in train and *all* roles in development and test speak for less than 5 minutes on average. This should be compared with other corpora used in prior lexical speaker ID studies: the target set in [11] is mostly restricted to speakers who spoke for at least 10 sessions i.e. 50 mins, with 9 sessions used for training. In [27] and [18], Fisher and NIST 1-side (5 mins) and 8-side (40 mins) data was used. In [4], target set is restricted to speakers who recorded at least 5 sessions, with a median duration of 6:30 min per session. In our experiments, no filtering based on train or test duration was made in the evaluation. Table 3 lists the relative duration of the speaker roles R1-R2 in each show S1-S7. Note dominant roles in bold.

The official REPERE protocol is followed in this work. The training corpus is used to build target speaker models, development corpus to tune system parameters and test corpus to evaluate the tuned system. Manual annotations were permitted for training but not for development and test. Performance on test is evaluated in terms of Identification Error Rate (IER), the percentage of the total annotated duration of a show when the hypothesized speaker identity i^{HYP} did not match the reference speaker identity i^{REF}:

$$\text{IER} = \frac{1}{|T_A|} \int_{t \in T_A} \mathbf{1}_{\{i^{\text{HYP}}(t) \neq i^{\text{REF}}(t)\}} \, \mathrm{d}t \; \times 100\% \tag{9}$$

Here, T_A represents the manually annotated segments of total duration $|T_A|$ and $i^{\text{HYP}}(t)$ is found via Eq. 8. The reference speaker $i^{\text{REF}}(t)$ may be `NONE`

	% of total time by speaker role				
TV Show	R1	R2	R3	R4	R5
S1	25.2	14.2	11.5	**38.1**	11.0
S2	14.0	0.0	**51.5**	0.1	34.3
S3	28.7	0.0	3.4	**63.2**	4.6
S4	25.7	**74.3**	0.0	0.0	0.0
S5	**27.2**	2.5	20.8	**27.3**	22.2
S6	25.2	0.0	0.0	**53.0**	21.8
S7	2.1	2.0	0.0	4.1	**91.9**

Table 3 Relative durations (%) of different speaker roles for each TV show in REPERE database. In each row, the dominant role (by duration) is marked in **bold**. Please see Sec.4.1 for details.

when there is no speech at time t (4.5% of the time) making the IER sensitive to speech/non-speech segmentation errors. Also, it may refer to more than one speaker in the case of overlapped speech (3.8% of the time). Again, this leads to errors because the current system treats overlapped segments in the same way as other segments and always hypothesizes one speaker at each time t.

Two forms of IER are reported in this work.

- **Open set IER** It is calculated using Eq. 9. The number of target speaker models is 359 and number of reference speakers in test is 132. We report open set IER for the main experimental results because it closely follows the guidelines of the official REPERE protocol. However, open set IER cannot distinguish errors due to speakers *not in* the training corpus from confusion errors between speakers *in* the training corpus.
- **Closed set IER** It is calculated using Eq. 9 but considering only those time instants t in test for which the reference speaker $i^{\mathrm{REF}}(t)$ exists in the training corpus. This means that the number of target speaker models is still 359 while the number of reference speakers in test is reduced to 63 *i.e.* the common speaker set.
  Computed in this way, closed set IER focuses only on the confusion between speakers in the training corpus and hence gives a better idea of the discriminative power of the SID system. Most of the analyses reported in Section 4.3 are based on the closed set IER.[6]

4.2 System description and notation

Independent SID systems using the four lexical approaches in Sections 3.1, 3.2, 3.3 and 3.4 were tested, denoted by TFIDF, BM25, MTW and LDA respectively. These were compared with a standard cepstral-based GSV-SVM

[6] Note that all 359 target speaker models are retained while scoring and not just the 63 in common so that the level of difficulty (equivalently the random chance performance) is at the same level as the open set case.

	System components			
	Training		Testing	
Configuration	Transcription	Diarization	Transcription	Diarization
mmMM	M	M	M	M
mmMA	M	M	M	A
mmAM	M	M	A	M
mmAA	M	M	A	A
amAA	A	M	A	A

Table 4 Brief overview of all system configurations studied. Columns 2-5 represents components of the system implemented either manually (M) or automatically (A). More details in Section 4.2.

SID system using supervectors made by concatenating UBM-adapted GMM means to train one SVM classifier per speaker as described in [22].

For each lexical system, five configurations were studied as shown in Table 4. In the first four configurations (mmMM, mmMA, mmAM and mmAA) manual transcription and diarization were used for *training* as permitted by the REPERE protocol (ref. Section 4.1) while for *testing*, four configurations were studied. 1) mmMM: manual transcription and diarization. 2) mmAM: automatic transcription and manual diarization, 3) mmMA: manual transcription and automatic diarization, and 4) mmAA: automatic transcription and automatic diarization. The first configuration (mmMM) shows the ideal scenario of having error-free transcription and diarization and sets the upper bound for performance. The fourth configuration (mmAA) is nearer to a practical scenario and follows the REPERE protocol. Intermediate configurations mmAM and mmMA are for analysis.

The fifth configuration, amAA, is the same as the fourth, mmAA, with the difference: it uses *automatic* transcripts for training instead of manual ones. Note that the amount of manual annotation involved in each configuration reduces as we go from the top row (mmMM) to the bottom row (amAA) in Table 4.[7]

4.3 Results and discussions

Tables 5, 6, 7 and 8 show the closed set and open set IER on REPERE test corpus for each show S1-S7 and time-averaged over all shows. The IERs are computed using GSV-SVM and lexical SID systems in configurations mmMM and mmAA. For the open set case *i.e.* Tables 6 and 8, we additionally provide the Oracle IER which measures the error due to speakers in test not existing in the training corpus and sets the open set IER lower bound.

[7] It is assumed that manual transcription takes more effort than manual diarization/segmentation.

	Closed set IER				
		Lexical (mmMM)			
Show	**GSV-SVM**	TFIDF	BM25	MTW	LDA
S1	24.2	31.8	44.8	44.8	58.7
S2	48.3	**34.7**	**26.4**	**35.4**	**35.8**
S3	20.0	32.3	32.3	**15.7**	43.4
S4	17.5	60.1	**0.0**	20.2	66.7
S5	22.2	39.1	44.9	43.3	59.2
S6	33.8	71.7	50.6	71.7	78.3
S7	11.3	58.3	58.3	58.3	94.1
All shows	26.8	46.2	39.5	43.6	65.4

Table 5 Closed set IERs of GSV-SVM and lexical systems in mmMM configuration on REPERE test corpus for shows S1-S7 (row 1-7) and over all shows (row 8). For each show (row), **bold** indicates a lexical system IER *lower* than corresponding GSV-SVM IER in column 3. Underlines mark the lowest IER among the 4 lexical systems.

Among lexical systems, proposed BM25 system and MTW system perform best, TFIDF next and LDA worst, for both mmMM and mmAA. Lower overall IER for BM25 than TFIDF for both mmMM and mmAA configurations indicate the utility of the nonlinear transformation in Eq. 3 as opposed to the TFIDF formulation in Eq. 2. Topic modeling performs poorly probably due to limited amount of training data in REPERE (around 300K words).[8]

Note that for mmMM configuration, BM25 and MTW outperforms GSV-SVM acoustic system in 2 out of 7 shows each while TFIDF and LDA outperforms GSV-SVM in 1 out of 7 shows each. For mmAA, only BM25 outperforms GSV-SVM in 1 out of 7 shows (S4). However, for show S2, MTW at 51.1% is close behind 48.3% achieved by GSV-SVM in Table 7.

Tables 9 and 10 break down the closed set IERs on REPERE test corpus in terms of speaker roles for GSV-SVM and lexical systems in mmMM and mmAA configurations respectively. Note that role information was used only to extract the test segments associated with a role while calculating IER, not to filter out speaker models with other roles.

The performance of BM25 and MTW systems is particularly good for R1 and R2 (which together comprise 31.1% of test time) and compares well with the acoustic system. TFIDF, BM25 and MTW perform moderately for R3, do not work for R4, and perform poorly for R5. This may be due to more training data for R1 and R2 (Table 2). However, R5 performs better than R4, although average training duration for R5 is lower than R4. This shows that factors other than duration may also be present, such as linguistic factors. This will be studied later. Except for a few cases, LDA performs poorly, compared the other three lexical approaches.

[8] Only REPERE training corpus was used to train topics so as to have perfectly matching training data. The option of using other corpora for training topics will be studied later.

	Open set IER					
			Lexical (mmMM)			
Show	**Oracle**	**GSV-SVM**	TFIDF	BM25	MTW	LDA
S1	49.6	61.8	65.6	72.2	72.2	79.2
S2	28.3	62.9	**53.2**	**47.2**	**53.7**	**54.0**
S3	41.5	53.2	60.4	60.4	**50.7**	66.9
S4	2.1	19.2	60.9	**2.1**	21.9	67.4
S5	43.2	55.8	65.4	68.7	67.8	76.8
S6	32.8	55.5	81.0	66.8	81.0	85.4
S7	14.8	24.4	64.5	64.5	64.5	95.0
All shows	35.5	52.8	65.3	61.0	63.6	77.7

Table 6 Open set IERs of Oracle, GSV-SVM and lexical systems in mmMM configuration on REPERE test corpus for shows S1-S7 (row 1-7) and over all shows (row 8). For each show (row), **bold** indicates a lexical system IER *lower* than corresponding GSV-SVM IER in column 3. Underlines mark the lowest IER among the 4 lexical systems. Oracle indicates open set IER lower bound.

	Closed set IER				
		Lexical (mmAA)			
Show	**GSV-SVM**	TFIDF	BM25	MTW	LDA
S1	24.2	73.6	54.2	45.6	92.9
S2	48.3	60.4	58.0	51.1	60.0
S3	20.0	39.2	38.6	39.2	48.0
S4	17.5	68.6	**15.8**	18.2	71.3
S5	22.2	55.3	55.1	58.3	93.0
S6	33.8	82.3	82.3	82.3	100.0
S7	11.3	67.4	57.8	57.8	84.7
All shows	26.9	66.2	51.9	49.9	83.0

Table 7 Closed set IERs of GSV-SVM and lexical systems in mmAA configuration on REPERE test corpus for shows S1-S7 (row 1-7) and over all shows (row 8). For each show (row), **bold** indicates a lexical system IER *lower* than corresponding GSV-SVM IER in column 3. Underlines mark the lowest IER among the 4 lexical systems.

Table 11 shows the closed set IER on REPERE test corpus of the lexical systems for the 3 configurations mmMM, mmAM and mmAM, to compare the impact of errors due to automatic transcription and automatic diarization in the testing phase individually. In the training phase, manual transcription and diarization were used for *all* 3 configurations, hence the prefix 'mm' is dropped from the table for easier reading. Note the difference in IER as we go from MM to AM (column 5) by replacing manual transcription with automatic transcription, and from MM to MA (column 6) by replacing manual diarization with automatic diarization in testing. It is observed that automatic diarization degrades IER much more than automatic transcription, for TFIDF, BM25 and

	Open set IER					
			Lexical (mmAA)			
Show	**Oracle**	**GSV-SVM**	TFIDF	BM25	MTW	LDA
S1	49.6	61.8	86.7	76.9	72.6	96.4
S2	28.3	62.9	71.6	69.9	64.9	71.3
S3	41.5	53.2	64.4	64.1	64.4	69.6
S4	2.1	19.2	69.3	**17.6**	19.9	71.9
S5	43.2	55.8	74.6	74.5	76.3	96.0
S6	32.8	55.5	88.1	88.1	88.1	100.0
S7	14.8	24.4	72.2	64.0	64.0	87.0
All shows	35.5	52.8	78.2	69.0	67.7	89.0

Table 8 Open set IERs of Oracle, GSV-SVM and lexical systems in mmAA configuration on REPERE test corpus for shows S1-S7 (row 1-7) and over all shows (row 8). For each show (row), **bold** indicates a lexical system IER *lower* than corresponding GSV-SVM IER in column 3. Underlines mark the lowest IER among the 4 lexical systems. Oracle indicates open set IER lower bound.

	Role-specific IER				
		Lexical (mmMM)			
Role	**GSV-SVM**	TFIDF	BM25	MTW	LDA
R1	12.8	**3.14**	13.3	13.7	21.6
R2	13.3	51.6	**0.0**	17.5	73.4
R3	13.9	27.0	34.4	34.4	57.7
R4	37.3	88.0	79.8	82.8	100.0
R5	25.4	70.2	67.0	69.2	97.4

Table 9 Closed set IERs of GSV-SVM and lexical systems in mmMM configuration on REPERE test corpus divided into speaker roles. For each role, **bold** indicates a lexical IER lower than GSV-SVM IER in column 3 and underlines show the lowest IER among 4 lexical systems.

	Role-specific IER				
		Lexical (mmAA)			
Role	**GSV-SVM**	TFIDF	BM25	MTW	LDA
R1	12.8	41.8	25.7	16.4	84.3
R2	13.3	56.3	**11.1**	13.1	55.8
R3	13.9	44.1	46.5	50.6	48.1
R4	37.3	98.5	97.0	96.4	99.7
R5	25.4	73.0	63.8	64.8	83.9

Table 10 Closed set IERs of GSV-SVM and lexical systems in mmAA configuration on REPERE test corpus divided into speaker roles. For each role, **bold** indicates a lexical IER lower than GSV-SVM IER in column 3 and underlines show the lowest IER among 4 lexical systems.

	IER			ΔIER	
System	MM	AM	MA	MM → AM	MM → MA
TFIDF	46.2	46.2	62.0	0.0	+15.8
BM25	39.5	45.0	54.0	+ 5.5	+ 14.5
MTW	43.6	42.6	60.3	-1.0	+ 16.7
LDA	65.4	83.9	73.6	+ 18.4	+ 8.2

Table 11 Relative impact of automatic transcription and diarization in testing phase, in terms of change in closed set IER on REPERE test corpus. In configuration names MM, AM, etc, first letter denotes transcription, second denotes diarization, in the testing phase. So, MA denotes **M**anual transcription and **A**utomatic diarization. Note that training always involved manual transcription and diarization in this specific study (mm).

	IER		ΔIER
System	mm	am	mm → am
TFIDF	66.2	52.2	-14.0
BM25	51.9	53.5	1.6
MTW	49.9	50.5	0.6
LDA	82.9	82.6	-0.3

Table 12 Impact of replacing *manual* transcriptions (mm) by *automatic* transcriptions (am) in training, in terms of change in closed set IER on REPERE test corpus. Diarization was always manual for training. For testing, both transcription and diarization was automatic (AA).

MTW. This shows that it would be more profitable to improve the diarization module rather than the ASR module in future for these 3 systems. This trend is reversed for LDA which is shown to be more sensitive to the impact of automatic transcriptions.

Table 12 shows the closed set IER on REPERE test corpus of the lexical systems for the 2 configurations mmAA and amAA, *i.e.* it shows the impact of replacing manual transcriptions in the training phase by automatic transcriptions. In this study, automatic transcription and diarization was always used for testing, so the suffix 'AA' was dropped in the table for easier reading. Note that when we go from mm to am, IER increases only slightly for BM25 and MTW, while for LDA it decreases slightly. Interestingly, TFIDF IER decreases significantly due to this change. Overall, this shows that it is possible for lexical SID systems to use automatic transcriptions for training rather than depending on manual annotation, without significantly affecting performance.

Table 13 shows the impact of duration of training data τ per speaker on closed set IER using the acoustic GSV-SVM system and the lexical BM25 system. Three conditions were studied: 1) $\tau \leq 120s$, 2) $120s < \tau \leq 420s$,

	IER for different training durations τ		
System	$\tau \leq 120s$	$120s < \tau \leq 420s$	$\tau > 420s$
GSV-SVM	36.6	12.5	14.1
BM25-mmMM	82.3	38.4	21.3
BM25-mmAA	82.4	44.4	27.5

Table 13 Impact of duration of training data τ per speaker (measured in seconds s) on closed set IER on REPERE test corpus using GSV-SVM acoustic system (row 1) and BM25 lexical system in manual (mmMM) and automatic (mmAA) testing configurations.

and 3) $\tau > 420s$.[9] Note that for durations less than 2 minutes, the BM25 system can still identify speakers approximately 18% of the time (IER ≈ 82). As the duration is increased, BM25 IER continues to decrease steadily. In contrast, although GSV-SVM system is always better than BM25, GSV-SVM IER saturates after training duration τ increases beyond 7 minutes.

4.4 Fusion studies

Acoustic- and lexical approaches rely on complementary sources of information to perform speaker identification. So, the latter is expected to improve the performance of the former when combined with it. In this paper, we report preliminary fusion studies based on simple sum fusion. Let $\mathrm{S}^{\mathrm{A}}(i,j)$ and $\mathrm{S}^{\mathrm{L}}(i,j)$ represent scores for a test speaker cluster j against target speaker model i, using the Acoustic GSV-SVM system and one of the Lexical systems respectively.[10] Then, the sum fusion scores are computed as:

$$S^{\mathrm{A+L}}(i,j) = B^{\mathrm{A}}(\mathrm{S}^{\mathrm{A}}(i,j)) + B^{\mathrm{L}}(\mathrm{S}^{\mathrm{L}}(i,j)) \tag{10}$$

respectively. Here, B^{A} and B^{L} are calibration functions to map the scores from the two different systems to posterior probability estimates. They are defined as:

$$B(\mathrm{S}(i,j)) = \frac{P_i \cdot \exp \rho(\mathrm{S}(i,j))}{\sum_{i':\mathrm{C}_{i'} \in \mathrm{C}_{\mathrm{tr}}} P_{i'} \cdot \exp \rho(\mathrm{S}(i',j)) + P_u} \tag{11}$$

where $P_u \approx 0.35$ is the prior probability for unknown speakers (*i.e.* speakers not present in training corpus), $P_i = (1 - P_u)/|\mathrm{C}_{\mathrm{tr}}|$ (same prior for every target speaker i) and $\rho(\mathrm{S}(i,j)) = \log \frac{p(\mathrm{S}(i,j)|i^{\mathrm{REF}}=i)}{p(\mathrm{S}(i,j)|i^{\mathrm{RFEF}}\neq i)}$ is the log-likelihood ratio estimated by linear regression on the development corpus. The fusion scores are then used in the same way as in Section 3.5.

[9] In this study, the set of 63 speakers in common between training and test was used both for training target speaker models and as the reference set in test (ref. Section 4.1). The time limits of 120 s and 420 s were chosen as round numbers which divide this set nearly equally into 20 speakers for each of the 3 conditions (precisely, 20, 20 and 23 speakers respectively).

[10] The scores for the GSV-SVM system were the distances of the test points to the decision hyperplane.

System	Overall IER	IER for each speaker role				
		R1	R2	R3	R4	R5
GSV-SVM	23.9	12.8	13.3	13.9	37.3	25.0
GSV-SVM + BM25-mmAA	**23.4**	**12.0**	**12.1**	13.9	37.3	25.0
GSV-SVM + BM25-mmMM	**21.4**	**6.7**	**8.5**	**6.1**	**34.6**	28.9

Table 14 Overall and role-specific closed set IER on REPERE test corpus using GSV-SVM acoustic system (row 1) and its fusion with BM25 in manual (mmMM) and automatic (mmAA) testing configurations. Bold indicates a fusion IER *lower* than corresponding GSV-SVM IER.

Table 14 shows the closed set IER on REPERE test corpus using the score-level sum fusion between the acoustic GSV-SVM system and the lexical BM25 system in manual (mmMM) and automatic (aaAA) testing configurations. For comparison, the IER for GSV-SVM system alone is shown in the first row. Overall, BM25 can reduce the acoustic system IER by 2.5% and 0.5% in manual and automatic configurations respectively (column 2). Also, BM25 improves the acoustic system IER for speaker roles R1-R4 individually in both manual and automatic configurations. The other lexical approaches studied did not show significant improvement on fusion.

5 Crossmedia lexical speaker identification

In previous sections, we used speaker-specific information extracted from speech transcripts for lexical SID. However, one may argue that we can always use standard cepstral-based systems in such settings and obtain better results, because whenever we have transcripts, we also have speech (or had speech at some point in the processing chain before it was transcribed). So, what is the value of lexical approaches other than improving cepstral-based system performance upon fusion?

In this section, we propose a new use case for SID which brings out the value of lexical approaches: Instead of using acoustic data for training cepstral-based speaker models or lexical content of speech derived from speech transcripts, this approach extracts speaker- or person-specific information from freely available Internet documents to train speaker or person lexical models. The resulting speaker or person models may be tested on ASR transcripts in the same way as in previous sections. This brings out the value of purely lexical approaches because it does not require any acoustic data for training. Since the training is on Wikipedia texts while the testing is on speech transcripts, we term this approach as *crossmedia*.

First, Wikipedia articles[11] about all speakers in the REPERE corpus were automatically searched, downloaded, parsed and normalized. Out of 474 target

[11] Only biographical articles were considered in this work, *i.e.* one article per speaker, *e.g.* `http://fr.wikipedia.org/wiki/Luc_Besson`.

speakers in the corpus, 330 had Wikipedia articles. These 330 speakers formed the speaker set for this study and their normalized Wikipedia texts were used to build lexical speaker models in exactly the same way as the documents derived from speaker clusters in Section 3.

While testing, a simple protocol was followed. A hundred speakers were chosen at random out of 330 and their Wikipedia-based lexical speaker models were matched with their speech transcripts in terms of their BM25 similarity score (ref. Section 3.2).[12] For each transcript, the speaker whose Wikipedia-based model obtained the highest BM25 score when matched with the transcript was chosen as the hypothesized speaker for the transcript.

Table 15 shows the performance of the crossmedia SID system in terms of 1-best, 2-best, 10-best and 20-best identification accuracies averaged over 20 independent runs, each run with 100 randomly chosen speakers. Both role-specific accuracies (rows 1-5) and overall accuracies (row 6) are shown. Random chance values are provided for comparison in row 7. Although the overall 1-best accuracy is low at 11.9%, the 10- and 20-best accuracies are 43.9 and 61.8% *i.e.* the system could retrieve the true speaker in the top 10 and top 20 shortlists 43.9% and 61.8% of the time respectively, significantly better than random chance. This trend is repeated for the role-specific cases too, except for role R2. This brings out the potential of this approach.

Among different roles, R1 (anchors) and R3 (reporters) have higher 1-best accuracies than others. As a possible explanation, it was found that R1 and R3 speakers typically mention the name of the associated TV channel or show and these names are also mentioned in their Wikipedia articles.

The accuracy for R4 (guests) increases steadily with the length of the shortlist and the 20-best accuracy for R4 is in fact the highest over all roles at 70.8%. This may be due to the fact that guests often speak about a specific domain (politics, movies, music, etc), which also shows up in their Wikipedia articles. The system is unable to deal with R2 (journalists). As a possible explanation, it was found that R2 speakers typically do *not* mention the name of any TV channel or show, nor do they speak about a specific domain.

Comparing Table 15 with Tables 9 and 10, it is observed that transcript-based lexical SID systems show complementary behaviour compared to crossmedia systems, performing best on R2 and worst on R4. This shows the potential of combining the two approaches.

6 Conclusions and future work

This paper reports four lexical speaker identification systems, one of which (BM25) has not been applied before on this task. Experiments were conducted using unsegmented TV shows from the REPERE database. Consistent results are reported for a wide spectrum of experimental conditions. Lexical approaches, specially BM25 and MTW, perform well for certain TV shows and

[12] Only manual transcripts were used in this study.

	Identification accuracy (%)			
Role	1-best	2-best	10-best	20-best
R1	38.4	46.5	54.0	62.6
R2	0.0	0.0	0.0	0.0
R3	23.6	33.6	45.1	47.3
R4	15.1	22.3	55.0	70.8
R5	9.7	14.9	40.4	59.9
All roles	11.9	17.7	43.9	61.8
Random chance	1.0	2.0	10.0	20.0

Table 15 Crossmedia lexical SID accuracy (%) using Wikipedia texts for training and speech transcripts from REPERE corpus for testing. The results are averaged over 20 independent runs with 100 test speakers uniformly drawn out of 330 speakers in each run. SID accuracy is shown for each role (rows 1-5) as well as over all roles (row 6).

speaker roles, sometimes better than the state-of-the-art GSV-SVM acoustic system. Upon fusion with the acoustic system, BM25 also succeeds in reducing the IER in both manual and automatic configurations, showing the potential of this approach. Interestingly, for most of the lexical systems, errors due to automatic diarization has a larger impact on IER than automatic transcription. Finally, an initial study on crossmedia SID showed promising results.

Motivated by the findings in this paper, priority will be given to improving the diarization system in future. We shall also explore hybrid acoustic-lexical approaches such as duration conditioned word n-grams previously found to out-perform purely lexical approaches [18,27]. We shall also extend the system to use multigrams instead of unigrams and transcribed text from previous and following speaker turns in addition to the text spoken by the speaker to be identified. The crossmedia SID system shall also be improved by using other text sources such as news websites.

There could be several applications of this work. The main application is the use of lexical information to improve the performance of speaker identification systems as shown here. The idea of using lexical information could also be extended to other similar tasks *e.g* detection of speakers in terms of profession, educational background, ethnicity and role. The crossmedia SID system shown here could be further developed to provide initial (possibly ambiguous) speaker labels to unannotated speech data for a semi-supervised acoustic SID system, reducing the need of costly human annotation.

Acknowledgements The authors would like to thank Lori Lamel for providing the alignments for the mmAM configuration, and François Yvon, Sophie Rosset, Sylvain Meignier and the anonymous reviewers for their helpful comments and advice. This work was partly realized as part of the Quaero Program and the QCompere project, respectively funded by OSEO (French State agency for innovation) and ANR (French national research agency).

References

1. Alam, M.J., Kinnunen, T., Kenny, P., Ouellet, P., O'Shaughnessy, D.: Multitaper MFCC and PLP features for speaker verification using i-vectors. Speech Commun. **55**(2), 237–251 (2013)
2. Baker, B., Vogt, R., Mason, M., Sridharan, S.: Improved Phonetic and Lexical Speaker Recognition through MAP Adaptation. In: Proc. of Odyssey (2004)
3. Barras, C., Zhu, X., Meignier, S., Gauvain, J.L.: Multi-stage speaker diarization of broadcast news. IEEE Transactions on Audio, Speech and Language Processing **14**(5), 1505–1512 (2006)
4. Baum, D.: Recognising speakers from the topics they talk about. Speech Communication **54**, 1132–1142 (2012)
5. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
6. Campbell, W., Campbell, J., Reynolds, D., Jones, D., Leek, T.: Phonetic Speaker Recognition with Support Vector Machines. In: Proc. Neural Information Processing Systems Conference, pp. 1377–1384 (2003)
7. Campbell, W., Sturim, D., Reynolds, D.: Support Vector Machines using GMM Supervectors for Speaker Verification. IEEE Signal Processing Letters **13**(5), 308–311 (2006)
8. Canseco, L., Lamel, L., , Gauvain, J.L.: A Comparative Study Using Manual and Automatic Transcriptions for Diarization. In: Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) (2005)
9. Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P., Dumouchel, P.: Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In: Proc. of Interspeech, pp. 1559–1562 (2009)
10. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-End Factor Analysis for Speaker Verification. IEEE Transactions on Audio, Speech and Language Processing **19**(4), 788–798 (2011)
11. Doddington, G.: Some experiments on ideolectal differences among speakers. Tech. rep. (2001). http://www.nist.gov/speech/tests/spk/2001/doc/
12. Doddington, G.: Speaker recognition based on idiolectal differences between speakers. In: Proc. of Interspeech, pp. 2521–2524 (2001)
13. Galibert, O., Kahn, J.: The First Official REPERE Evaluation. In: Proc. of Workshop on Speech, Language and Audio in Multimedia (SLAM) (2013)
14. Gauvain, J.L., Lamel, L., Adda, G.: Partitioning and transcription of broadcast news data. In: Proc. of International Conference on Spoken Language Processing (ICSLP), pp. 5:1335–1338 (1998)
15. Gauvain, J.L., Lamel, L., Barras, C., Adda, G., de Kercadio, Y.: The LIMSI SDR System for TREC-9. In: Proc. of TREC-9 (2000)
16. Giraudel, A., Carre, M., Mapelli, V., Kahn, J., Galibert, O., Quintard, L.: The REPERE corpus : a multimodal corpus for person recognition. In: Proc. of Language Resources and Evaluation Conference (LREC) (2012)
17. Jones, K.S., Walker, S., Robertson, S.: A probabilistic model of information retrieval: Development and comparative experiments. Information Processing and Management **36**(6), 779–840 (2000)
18. Kajarekar, S.S., Ferrer, L., Shriberg, E., Sonmez, K., Stolcke, A., Venkataraman, A., Zheng, J.: SRI's 2004 NIST Speaker Recognition Evaluation System. In: Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2005)
19. Khan, A., Yegnanarayana, B.: Latent semantic analysis for speaker recognition. In: Proc. of International Conference on Spoken Language Processing (ICSLP) (2004)
20. Kinnunen, T., Li, H.: An Overview of Text-Independent Speaker Recognition: From Features to Supervectors. Speech Communication **52**(1), 12–40 (2010)
21. Lamel, L., Courcinous, S., Despres, J., Gauvain, J.L., Josse, Y., Kilgour, K., Kraft, F., Le, V.B., Nussbaum-Thom, H.N.M., Oparin, I., Schlippe, T., Schluter, R., Schultz, T., da Silva, T.F., Stuker, S., Sundermeyer, M., Vieru, B., Vu, N.T., Waibel, A., Woehrling, C.: Speech recognition for machine translation in quaero. In: Proc. of IWSLT (2011)
22. Le, V., Barras, C., Ferras, M.: On the use of GSV-SVM for speaker diarization and tracking. In: Proc. of Odyssey, pp. 146–150 (2010)

23. Manning, C., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
24. Mauclair, J., Meignier, S., Esteve, Y.: Speaker diarization: About whom the speaker is talking? In: Proc. of Odyssey (2006)
25. McCallum, A.: Mallet: A machine learning for language toolkit. Tech. rep. (2002). http://mallet.cs.umass.edu
26. Plchot, O., Matsoukas, S., Matejka, P., Dehak, N., Ma, J., Cumani, S., Glembek, O., Hermansky, H., Mallidi, S., Mesgarani, N., Schwartz, R., Soufifar, M., Tan, Z., Thomas, S., Zhang, B., Zhou, X.: Developing a Speaker Identification System for the DARPA RATS Project. In: Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2013)
27. Shriberg, E.: Higher-level features in speaker recognition. Speaker Classification I, LNAI 4343 pp. 241–259 (2007)
28. Tran, V.A., Le, V., Barras, C., Lamel, L.: Comparing multi-stage approaches for cross-show speaker diarization. In: Proc. of Interspeech, pp. 1053–1056 (2011)
29. Tranter, S.: Who really spoke when? Finding speaker turns and identities in broadcast news audio. In: Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2006)
30. Tur, G., Shriberg, E., Stolcke, A., Kajarekar, S.: Duration and Pronunciation Conditioned Lexical Modeling for Speaker Verification. In: Proc. of Interspeech (2007)