Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives

William Byrne, *Member, IEEE*, David Doermann, Martin Franz, *Senior Member, IEEE*, Samuel Gustman, Jan Hajič, Douglas Oard, Michael Picheny, *Fellow, IEEE*, Josef Psutka, Bhuvana Ramabhadran, *Member, IEEE*, Dagobert Soergel, Todd Ward, and Wei-Jing Zhu

***Abstract*—Much is known about the design of automated systems to search broadcast news, but it has only recently become possible to apply similar techniques to large collections of spontaneous speech. This paper presents initial results from experiments with speech recognition, topic segmentation, topic categorization, and named entity detection using a large collection of recorded oral histories. The work leverages a massive manual annotation effort on 10 000 h of spontaneous speech to evaluate the degree to which automatic speech recognition (ASR)-based segmentation and categorization techniques can be adapted to approximate decisions made by human annotators. ASR word error rates near 40% were achieved for both English and Czech for heavily accented, emotional and elderly spontaneous speech based on 65–84 h of transcribed speech. Topical segmentation based on shifts in the recognized English vocabulary resulted in 80% agreement with manually annotated boundary positions at a 0.35 false alarm rate. Categorization was considerably more challenging, with a nearest-neighbor technique yielding $F = 0.3$. This is less than half the value obtained by the same technique on a standard newswire categorization benchmark, but replication on human-transcribed interviews showed that ASR errors explain little of that difference. The paper concludes with a description of how these capabilities could be used together to search large collections of recorded oral histories.**

***Index Terms*—Automatic speech recognition (ASR), information retrieval, multilingual ASR, oral history, spoken document retrieval, spontaneous speech.**

Manuscript received May 10, 2003; revised January 25, 2004. This work was supported in part by the National Science Foundation (NSF) IIS Award 0122466 (MALACH), by NSF CISE Research Infrastructure Award EIA0130422, by the Ministry of Education of the Czech Republic projects LN00A063, MSM235200004, and MSM113200006, by the Grant Agency of the Czech Republic Grant 405/03/0913, and by a Shared University Research grant from IBM. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Julia Hirschberg.

W. Byrne is with the Center for Language and Speech Processing and the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: byrne@jhu.edu).

D. Doermann is with the Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 USA (e-mail: doermann@umiacs.umd.edu).

M. Franz, T. Ward, and W.-J. Zhu are with the Natural Language Systems Department, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: franzm@us.ibm.com; toddward@us.ibm.com; weijing@us.ibm.com).

S. Gustman is with the Survivors of the Shoah Visual History Foundation, Los Angeles, CA 90078 USA (e-mail: sam@vhf.org).

J. Hajič is with the Institute of Formal and Applied Linguistics and Center for Computational Linguistics, Charles University, CZ-11800 Prague 1, Czech Republic (e-mail: hajic@ufal.mff.cuni.cz).

D. Oard is with the Institute for Advanced Computer Studies and College of Information Studies, University of Maryland, College Park, MD 20742 USA (e-mail: oard@glue.umd.edu).

M. Picheny and B. Ramabhadran are with the Human Language Technologies Department, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: bhuvana@us.ibm.com; picheny@us.ibm.com).

J. Psutka is with the Department of Cybernetics and the Center for Computational Linguistics, University of West Bohemia, CZ-30614 Pilsen, Czech Republic (e-mail: psutka@kky.zcu.cz).

D. Soergel is with the College of Information Studies, University of Maryland, College Park, MD 20742 USA (e-mail: dsoergel@umd.edu).

Digital Object Identifier 10.1109/TSA.2004.828702

I. Introduction

In a recent report, an international digital library working group called for the creation of systems capable of providing access to an estimated 100 million hours of culturally significant spoken word collections [1]. Achieving that bold vision will require two fundamental advances over the present state of the art: 1) a robust ability to identify spoken words and other useful features in many types of collections and 2) development of systems that can leverage those features to meet the real needs of real searchers. Recent work on searching collections of broadcast news (BN) indicates that these goals are now within the reach of our technology (e.g., [2]). There is, however, still a substantial gap between our vision and our grasp. In this paper, we focus one key implication of this broader challenge: the degree to which existing techniques can be adapted to provide access to spontaneous conversational speech.

Research on automated transcription of spontaneous conversational speech dates back to the creation of the Switchboard (SWB) corpus of telephone speech in 1993 [3]. Several telephone speech collections were developed over the next decade, and carefully tuned automatic speech recognition (ASR) systems are now able to achieve word error rates (WERs) between 20% and 40%, depending on the difficulty of the collection. In a series of annual evaluations that started in 1996, researchers participating in the Spoken Document Retrieval (SDR) track of the Text Retrieval Conferences (TREC) developed ranked-retrieval techniques for subject-based searching in BN collections that were robust in the presence of WERs in the 20%–40% range [4]. Similar results were obtained for exact-match event detection in BN at the Topic Detection and Tracking (TDT) evaluations. Research for the TDT evaluations also resulted in development of reliable detection of story boundaries in undifferentiated streams of BN speech, an important prerequisite to the story-based search techniques evaluated at TREC and TDT.

The next logical step in this evolutionary development path is to apply what we have learned from our work with BN to search large collections of spontaneous conversational speech. Unfortunately, none of the existing collections of telephone speech

that were created to support ASR research are suitable for this purpose. In some cases (e.g., SWB), callers were prompted to discuss a specific subject. In others (e.g., the CallHome series), speakers conversed with personal acquaintances; in those cases, substantive discussion of the same topic by multiple speakers turned out to be rare. We were, however, able to obtain access to an existing set of recordings that had been collected for another purpose that had exactly the characteristics that we needed for this research.

The Survivors of the Shoah Visual History Foundation (VHF), founded to preserve the stories of survivors and witnesses of the Holocaust [5], has created what we believe to be the largest collection of digitized oral history interviews on a single topic: almost 52 000 interviews in 32 languages, a total of 116 000 h of audio and video. About half of the collection is in English, and about 10 000 h of the English interviews have been extensively annotated by subject-matter experts with topic boundaries for approximately 200 000 segments (with an average duration of 3 min) and multiple topic labels for each segment. The annotation effort also resulted in careful transcription of all person names in each segment, tagging of each segment with the location and year, and the creation of brief (3-sentence) summaries for each segment. The annotator's personal ("scratchpad") notes for each segment, which include a variable degree of detail, were also retained. Topic and location labels are keyed to a thesaurus with multiple inheritance that encodes both part-whole and is-a relationships. This segment-level description is augmented by three data items associated with the interview as a whole: a name authority file that resolves references to the same person within a single interview, a half-page summary of the interview written by the annotator, and extensive demographic information provided by the interviewee as responses to a questionnaire. The annotation effort alone required approximately 150 000 h (75 person-years); collection and digitization of the interviews incurred additional expenses that are not included in this estimate.

This paper describes a series of experiments using the VHF collection that establish a strong foundation for research on search techniques for automatically transcribed spontaneous conversational speech. Section II describes in detail the approach that we are taking to build recognition systems for spontaneous speech, presenting results for English and Czech. Section III then describes three key natural language processing (NLP) components that use the output of these systems to achieve three fundamental capabilities that are needed to support a search process: named entity detection, topic segmentation, and segment categorization. Section IV then draws on these results, describing ways in which the capabilities demonstrated in this paper could be integrated to build effective systems for searching spontaneous conversational speech.

II. Automatic Spontaneous Speech Recognition

A. Goals and Challenges

We have two goals for speech recognition: 1) sufficient accuracy to support downstream processing such as search, boundary detection, and content annotation and 2) sufficient extensibility to support affordable application of similar techniques to incorporate additional languages or collections. Although WER is an imperfect measure for this application (because missing some words causes more problems than others), several studies with BN have shown little degradation for term-based techniques over a range of WERs, up to 40% [6]. Beyond that point, however, techniques based on counting term occurrences typically degrade rapidly. Our first goal is therefore to achieve a WER below 40% on as large a fraction of representative unseen test data as is possible. Our second goal, affordable extensibility, requires that we accomplish that with a minimum of condition-specific effort. Because the data that we are working with contains many languages but only a single domain of discourse, we have chosen a set of five languages (English, Czech, Russian, Polish, and Slovak) with which to explore this question.

In this section, we describe the characteristics of the spontaneous speech that we seek to recognize, our approaches to acoustic modeling and language modeling for English, Czech, and Russian, and the results we have obtained for English and Czech.

The VHF collection consists of unconstrained, natural speech filled with disfluencies, heavy accents, age-related coarticulations, un-cued speaker and language switching, and emotional speech. As will be described in Section II-B, transcription is challenging even for skilled annotators. Human transcribers typically required 8–12 h to transcribe a single hour of an English interview. The difficulties arise from unfamiliar names and places, multiple languages encountered during a single interview, coarticulations related to age, highly variable speaking rates, and heavily accented speech. For illustration, we include here an example of the actual words spoken during one segment:

> It wasn't everybody living in one in one one ghetto you know was a little like the in this street a was a house ghetto in this street it had ghetto but people couldn't people wasn't allowed to go out in the streets when they came in the Nazis came in he wanted they made a Jewish committee the Jewish committee have to help him take where to live and took out the furniture from from the from the Jewish people and so and Jewish committee had eighteen people with me also I helped the Jewish committee I mean the reason is they had eighteen people we walked the street everyday two two people two friends we walked on each street as people doesn't go out on the street ...

If the linguistic phenomena that must be modeled did not differ across languages, it would suffice to perform an in-depth study of a single language and then extend the results to the other languages. However, as described in Section II-C, very different phenomena arise in the three languages we have studied so far. The conservative approach to ASR is to collect as much in-domain data as possible to train the best performing system. However, for many practical applications, this approach is prohibitively expensive. If ASR is to be deployed for these other collections, the practical problem therefore is not simply "how much data is needed to train an ASR system?", but rather "what is the best systematic approach for inexpensive and fast development of systems in a new language?" Thus, our research aims

to develop ASR systems that are "good enough" for the uses to which they will be put, and to provide studies of "best practices" that can be used for future cost-effective ASR development for use in information retrieval systems.

Certain ASR problems, such as those that arise from elderly, spontaneous, and emotional speech, are common throughout the corpus. However, some other problems that we have observed are language-specific. For instance, the English language interviews tend to be highly accented because the majority were conducted with people who learned English as a second language. In contrast, the Russian speakers in the collection are predominantly native speakers. However, this does not mean that the Russian language interviews are any more uniform than the English interviews. There are regional effects on the speech collected within Russia, and most Russian interviews were conducted outside Russia, where the speakers were influenced by the languages of their adopted countries. Of the approximately 7000 Russian interviews in the collections, nearly half (3500) were taken in the Ukraine, 1500 in Israel, 900 in the United States, and only about 700 in Russia itself. As a practical problem in ASR, it may be adequate to model these Russian and English speakers using similar modeling techniques; however, the underlying causes of the variability—nonnative speech in English and regional or dialectal influences in Russian—are quite different.

There are also spontaneous speech effects that are entirely language-specific. Czech is spoken by a relatively small number of speakers within a relatively small geographic area, and it is therefore not surprising that the Czech language interviews are not particularly accented. However, spontaneously spoken Czech contains words and usages not found either in standard written or in formal spoken Czech. The VHF collection, in particular, is rich in these spontaneous forms. Examples are given in Table I, showing the differences between the colloquial usage that appears in our transcriptions and the formal versions that would appear if the same sentiment were to appear in news text or broadcast transcriptions. Many of these forms can be analyzed morphologically; however, usage is variable. As a result, automatically mapping formal text to its spontaneous form is problematic, because users are not consistent in their choice of spontaneous form. The best source of information on spontaneous usage is transcribed speech itself, which leads to a shortage of data for the construction of statistical language models.

Given the intended role of ASR to support information access, we are particularly interested in named entity recognition [7], [8] especially the recognition of personal names and place names which are both important search criteria (see Section IV). This is a difficult problem, since these names are often not in the ASR lexicon. Moreover, names come in many variations (Hebrew names, Yiddish names, diminutives, first names only, etc.). The VHF collection offers an opportunity to study this problem through its large database of personal identities (approximately 2.5 million names) that is populated with information taken from survey forms filled out by the subjects [known as pre-interview Questionnaires (PIQs)], additional names assigned by catalogers, and a large list of place names (over 20 000 locations). One important challenge in this work is that many key search terms will be found only among the infrequently occurring words and phrases, and rare terms are inevitably modeled less well than more common ones. Strategies to develop an optimal lexicon are addressed in Section II-D.

TABLE I
FORMAL CZECH AND COLLOQUIAL VARIATIONS

Formal	Oni mi opatřili český pas.
Colloquial	Voni mi vopatřili českej pas.
English	They provided a Czech passport for me.
Formal	... bývalý český romtistr
Colloquial	... bejvalej českej rotmistr
English	... a former Czech sergeant

B. Data Preparation

Approximately 25 000 of the collected interviews are in English, 7000 are in Russian, and 575 are in Czech. The average duration of an English interview is 2.5 h, the average length of a Czech interview is 1.9 h, and the average length of a Russian interview is 2 h. The interviews were recorded under a wide variety of conditions ranging from quiet to noisy (e.g., airplane overflights, wind noise, background conversations, and highway noise). Original interviews were recorded on Sony Beta SP tapes, then digitized into a 3 MB/s MPEG-1 stream with 128 kb/s (44 kHz) stereo audio.

The average speaking rate of the English interviewees is 146 words/min, with a dynamic range of 100–200 words/min. For comparison, the average speaking rate in the SWB corpus is 100 words/min [3]. The relatively high speaking rate in the English part of the VHF collection is not constant throughout the corpus, however. The average speaking rate in Czech is 113 words/min, with individual utterances ranging from 64 to 173 words/min; and the average speaking rate in Russian is 123 words/min; clearly speaking rate is highly variable.

In order to maximize the number of speakers in our training data, we chose to transcribe 15-min segments from a large number of interviews. For English, this 15-min passage was selected randomly from within an interview. For Czech and Russian, it always started 30 min into the interview, so as to skip the biographical material that was commonly discussed at the beginning of an interview.

The English corpus was generated using 15-min segments of an interview from 800 randomly selected speakers. Thus, a total of 200 h of data was selected for manual transcription that would subsequently serve as training material for ASR systems. Male and female speakers in this corpus were more or less equally distributed and a wide range of accents were covered (e.g., Hungarian, Italian, Yiddish, German, and Polish). The English ASR test set consists of 30-min segments taken from 30 randomly chosen speakers; results reported here are on a 1 h/20 speaker subset of this collection. An additional test set of five full interviews was also created to support evaluation of natural language processing (NLP) techniques (Section II-A).

The Czech corpus is smaller than the English corpus. The ASR training set consists of 84 h of speech taken from 336 speakers. The speakers are predominantly Bohemian in origin (73.4%), but there are also speakers from Slovakia (13.0%), and

the Carpathians (5.2%).[1] The Czech ASR test set consist of ten interviews transcribed from beginning to end; results reported here are on a 500-sentence subset of that collection.

The audio files were divided (roughly) into sentences by the human annotators. Transcription was done using the Transcriber 1.4.1 speech editing tool [9], which was modified to incorporate Unicode for the transcription of Russian and Czech. In addition to lexical transcription, the following nonspeech sounds were marked: Tongue click, lip smack, cough, laughter, breath, inhalation, UH, UM, background noise, silence, and unintelligible. Names, places and sections spoken in a different language were marked to the extent possible. The rules for the entire annotation process have been published in detail previously [10], [11]. The annotators worked at a rate of 15 times real time in Czech, 12 times real time in English and 18 times real time in Russian. Transcription inspection and verification requires additional effort at approximately twice real time. In contrast to other well-know transcription tasks, such as BN and SW conversations [3], the heavily accented and poorly articulated elderly speech posed severe problems for transcribers and this is reflected in Fig. 1.

The compressed audio signal in MP3 format was stored at a sampling frequency of 44.1 kHz. This signal was extracted from the video and down-sampled to 16 or 8 kHz to match the recognizer that was used. The original interviews were done in stereo with the interviewer and interviewee in separate channels. Given the nature of the collection effort, it was not possible to ensure that recording conditions were always ideal; for instance, it often happened that microphones were placed so that the speakers are equally loud in each channel, and it is not uncommon for both the interviewer and interviewee actually to have been recorded on the same channel. The occasional use of far-field microphones in a noisy environment introduces additional acoustic variability. A summary of the signal-to-noise-ratio (SNR) as distributed over the English interviews is given in Fig. 2. It can be seen that a significant fraction of the data is noisy, with an energy level below 10 dB.

While it is important to the task to be able to process both the speech of the interviewer and the interviewee, we decided to use only the channel on which the interviewee was the loudest, as the great majority of speech that occurs on either channel is from the interviewee; in this way we are assured of transcribing the greatest amount of speech for use as acoustic model training data.

C. Acoustic Modeling

Pronunciation lexicons were constructed for the collections based on existing English, Russian, and Czech resources. Pronunciations were added as needed by transcribers to cover the acoustic training set.

1) Acoustic Modeling—English: The audio signal was extracted from the video and down-sampled to 16 or 8 kHz to match the recognizer that was used. The 16-bit down-sampled PCM signal was used to produce 24-dimensional mel frequency cepstral coefficients (MFCC). The MFCC features

[1]Slovakia was part of the former Czechoslovakia from 1918 to 1992, and the Carpatho-Ukraine was part of the former Czechoslovakia from 1918 to 1938; people living there spoke Slovak and Ukrainian, respectively.

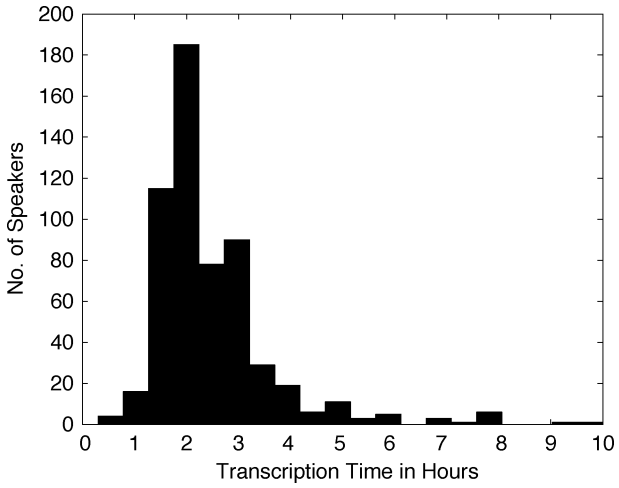


Fig. 1. Transcription times computed over the English acoustic training data.

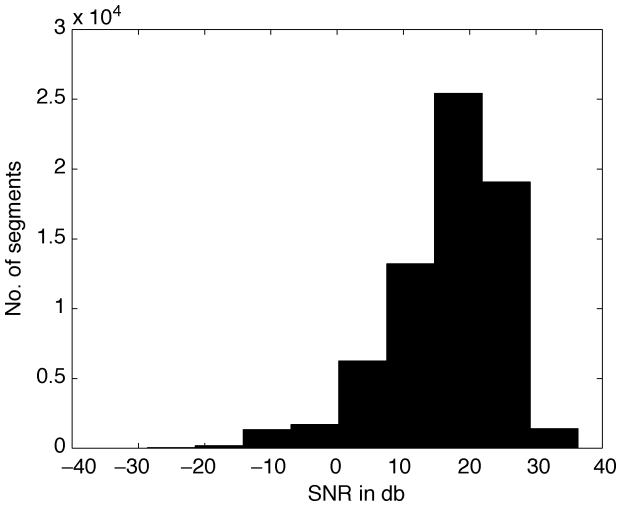


Fig. 2. SNR computed over the English acoustic training data.

were computed from a 24-filter Mel filterbank spanning the 0 Hz–4.0 kHz frequency range for the 8-kHz system (System I) and 0 Hz–8.0 kHz for the 16-kHz system (System II). These two bandwidths were selected to enable comparisons between state-of-the-art narrowband (SWB) and broadband (BN) systems respectively. All feature sets use 25-ms frames with a 10-ms step, perform spectral flooring by adding the equivalent of one bit of additive noise to the power spectra prior to Mel binning, and use periodogram averaging to smooth the power spectra. Every nine consecutive cepstral frames are spliced together and projected down to 60 dimensions using a linear discriminant feature space transformation to ensure maximum phonetic discriminability. The range of these transformations is further diagonalized by means of a maximum-likelihood linear transform (MLLT) [12] to decorrelate dimensions.

The transcriptions used initially for acoustic model training contained a significant number of transcription errors. Many of these errors were due to the many foreign words, names, places, and sequences of words uttered in a foreign languages (such as German, Yiddish, or Hebrew), that the transcribers found unfamiliar. The manual refinement of these transcriptions was aided by the use of the list of proper names provided by VHF; this

resource proved extremely valuable, since the transcribers did not have the specialized knowledge needed to identify words, places, and individuals particular to this domain. Initial model alignments were obtained using an existing ASR system built on SWB training data [12]. The average log-likelihood of each segment in the training data conditioned on the alignments was used to reject the segments that had either transcription errors or incorrect pronunciations in the lexicon. Pronunciations for the many unseen words in this corpus were derived with the help of existing dictionaries and tools using spelling-to-sound rules. These alignments were also used to build decision trees (DTs) [13] to capture the context-dependent variations of this speech and the data at the leaves of the DT were modeled with Gaussian distributions via a BIC-based procedure [12] and trained using multiple iterations of the EM algorithm. The final models for the two systems (System I and System II) had 3000 context-dependent states modeled by 60 K Gaussians.

Speaker-adaptive models (SAT) [14] were trained via a feature space MLLTs, i.e., for MLLR, for each training speaker. The canonical model was first initialized as the speaker-independent model. After fMLLR transforms for training speakers were computed against the canonical model, the canonical model was then re-estimated using the affinely transformed features. This method is based on the SAT principle, but differs slightly from SAT in that the normalization is applied to the features. This corresponds to using a constrained maximum-likelihood linear regressing (MLLR) [15] transform instead of a mean-only MLLR transform.

2) Acoustic Modeling—Czech: The acoustic training set consisted of approximately 84 h of speech. The data at 44.1 khz was parameterized as 15 dimensional PLP cepstral features, with first and second cepstral derivatives. Features were computed at a rate of 100 frames/s. Cepstral mean subtraction was applied per utterance. Cross-word triphone acoustic models were trained using the HTK Toolkit [16], [17]. The resulting models had approximately 6000 states and 96 K Gaussians. A silence model was trained by borrowing Gaussians from all the nonspeech HMMs in proportion to their state and mixture occupancies. The resulting model contained 176 mixtures per state and was found to be useful in rejecting nonspeech events during recognition.

D. Language Modeling

1) Language Modeling—English: Two basic language models were trained on transcriptions of both 65 h and 200 h of data from this corpus using the modified Kneser-Ney algorithm [18]. Since the training data is relatively small (320 k/1.7 M words, respectively), there is a high out-of-bocabulary (OOV) rate, especially among personal and place names. To address this, language models built from BN and SWB corpora (158 and 3.4 M words, respectively) were interpolated with the LM built from the VHF collection alone. The interpolated weights were optimized to achieve minimum perplexity on the held-out data from the VHF collection. The effect of an increase in the in-domain material and the interpolation across other speaking styles such as those seen in BN and SWB tasks are illustrated in Table II. With the addition of more in-domain data, the percentage of trigram counts used from the SWB and BN corpora decreased from 66%, when only 320 K in-domain words were used, to 26% when 1.7 M words were used. The average OOV rate on the test set was 8.2% (with a 3.2% to 11.6% range). The final lexicon consists of 64 k words derived from a combination of sources such as the PIQ, the cataloged segment data and the human transcriptions.

2) Language Modeling—Czech: This transcribed corpus (Tr, 84 h) is the only available spontaneous speech corpus for Czech language modeling. Given the limited size of this collection, investigations were performed to see whether the use of other, out-of-domain sources of text data would improve language model quality. Two collections were considered. The news text corpus Lidové Noviny (LN, 33 M words) was a convenient choice since it has been used in many other Czech LM experiments [10]. However initial experiments suggested that this collection was an inappropriate source of LM training data for this domain; in particular, the LN text was found to have a very high OOV rate (9.6%) on the task transcriptions. The Czech National Corpus (CNC) was also considered as a source of data. It is relatively large (approximately 400 M words), and is extremely diverse. Rather than use the entire corpus, we employed statistical methods to determine what portions of the CNC are similar in "colloquiality" and lexical choice to the interview transcriptions [19]. The hope was that, due to the size of the CNC, we might find a reasonable amount of data with high similarity. The method has been based on two simple unigram models: one, $P(\cdot|CNC)$, estimated over the entire CNC, and the other one, $P(\cdot|Tr)$, estimated over the transcriptions at hand. A likelihood ratio then determined whether a tested section of text was closer to the general CNC or to the transcribed spontaneous speech corpus (Tr). Varying the ratio threshold allowed us to effectively control the size of the corpus as has been done in previous work [20]. At a threshold of 0.8 the size of the selected corpus (named CNCS) was 15.8 M words. An interpolated model (Tr-CNCS) has also been created, showing the lowest OOV rate of all the models used (Tables IV and V).

E. ASR Performance

1) ASR Performance—English: We first present initial recognition results on the English portion of the VHF collection. A baseline WER was first computed using a speaker-independent, MFCC system, built for the SWB task [12]. This system was trained on the SWB, CallHome, CTIMIT, and the National Cellular Corpora. The acoustic model component consisted of 300 K diagonal covariance Gaussians, and the lexicon contained 64 K words. The system achieved an error rate of 47.3% on the spontaneous conversations in the SWB 1998 Evaluation task. However, the speaker-independent performance of this system was as high as 85.6% on this test data (Table III) compared to the baseline system trained on in-domain data. Similarly, the English system trained on VHF data alone (System II) has a performance of 69.4% on the SWB 1998 Evaluation task. It is interesting to note that even though both corpora are recorded conversations between two individuals, the drastic differences in the nature of the speech coupled

TABLE II
PERPLEXITY AND WORD ERROR RATES FOR VARIOUS ENGLISH LANGUAGE MODELS

LM Corpora	Perplexity	WER (%)
SWB + BN (LM0)	180	57.3
65 hours of VHF interpolated with SWB and BN (LM1)	95.1	54.1
200 hours of VHF	86.9	53.3
200 hours of VHF interpolated with SWB and BN (LM2)	72.3	53.1

with the high OOV rates result in poor speech-recognition performance.

The speaker-independent acoustic models from System I were used in the language modeling experiments results reported in Table II. It can be seen from Table II that both increasing the in-domain text data as well as incorporating data from other corpora produces a reduction in perplexity and the overall WERs. A gain of 4% absolute can be obtained when the in-domain data is tripled and augmented with similar data from other related corpora.

Table III presents the speech recognition results obtained using different acoustic models on this new task. The two systems used in these experiments were built by down-sampling the original audio signal recorded at 44.1 to 8 kHz (System I) and 16 kHz (System II). While both systems are comparable in performance, the wider bandwidth system (System II) has a relative 6% performance improvement over the bandlimited system. The speaker-independent system (System II) built on 65 h of VHF data produces a WER of 54.3% on this task (Row 2). This reiterates prior work in the literature that significant improvements, such as halving the WER, can be obtained when the acoustic models are trained using in-domain data. When this system is augmented with a language model that has been trained on the VHF collection, further improvements can be seen (Row 3). The SAT models reduce the error rate further to 43.6% (Row 4). Subsequent adaptation using MLLR and an improved language model results in a WER of 40.2% (Row 7). If transcripts at a lower WER (Row 5) obtained using techniques such as consensus decoding [12] are used for computing the MLLR transforms, the WER can be reduced further to 39.6% (Row 8).

As an initial investigation into refining the pronunciation lexicon for the task, we investigated the use of context-free syllabic models for use in a mixed syllabic-phonetic dictionary. Syllable clusters were derived using software available from NIST [21] and the lexicon used in Systems A and B. This software implements syllabification rules that define permitted syllable-initial consonant clusters, syllable-final consonant clusters and prohibited onsets [22]. For example, consider the word "BUDAPEST" which has "B UW D AX P EH S T" as one of its phonetic pronunciations. The English syllabic representation was "B UW D_AX P_EH_S T", and if enough data was available, acoustic models would be trained for each syllable. However the syllable "P_EH_S" occurs infrequently, so the pronunciation backs off to the phonetic level and the word has the "syllabic" pronunciation "B UW D_AX P EH S T". During recognition, a dictionary is created that has both the phonetic and syllabic forms for each word in parallel. The number of states in the syllable model were designed based on the number of phones in the syllable so that the syllable and phonetic model had the same complexity. The joint modeling of syllables and context-dependent phones provides a 0.5% absolute improvement in recognition accuracy, from 39.6% to 39.1% [23].

TABLE III
ENGLISH ASR PERFORMANCE (WER%)

System Description	System I (8 Khz)	System II (16 Khz)
SWB system + LM0	85.6	NA
Baseline on VHF + LM0	57.3	54.3
Baseline on VHF + LM1	54.15	51.3
SAT Baseline on VHF + LM1 (System A)	46.3	43.6
Consensus Decoding on System A (System B)	–	41.5
MLLR on System A	–	42.2
MLLR on System A + LM2	46.1	40.2
MLLR on System B + LM2	–	39.6

2) ASR Performance—Czech: The acoustic modeling and adaptation experiments reported here were performed using HTK with an interface developed for the AT&T Large Vocabulary Decoder [24]. Models and features generated with HTK are used by the AT&T Decoder to generate hypotheses for unsupervised MLLR adaptation and lattices for acoustic rescoring.

Results with each of the Czech language models are summarized in Table IV [19]. The effect of OOVs is readily apparent; the difference in performance between LM-Tr and LM-Tr-C is almost exactly the OOV rate with respect to the acoustic training set transcriptions. The worst performance is obtained under the Lidove–Noviny language model (LM-LN). We note that the degradation relative to LM-Tr is not explained by the LM-LN OOV rate alone; the LN corpus itself is clearly mismatched to this task. The language model trained on the filtered CNC (LM-CNCS) achieves performance that is worse than the one obtained with the LM-Tr model but significantly better than with the LM-LN.

Although the LM-CNCS language model did not perform as well as the models trained on the acoustic training set, we took advantage of their different vocabularies and created a merged model LM-Tr-CNCS: we merged the CNCS and Tr vocabularies, retrained bigram language models in each domain (obtaining LM-TrU and LM-CNCSU), and used the SRILM Toolkit [25] to interpolate them linearly using the following formula:

$$P_{\mathrm{LM-Tr-CNC-S}} = \lambda P_{\mathrm{LM-Tr'}} + (1-\lambda) P_{\mathrm{LM-CNC-S}}. \quad (1)$$

WERs are reported in Table V for some values of the parameter λ. For $\lambda = 0.0$, which corresponds to training on the CNCS corpus alone, we find an improvement relative to the 52.99% WER result reported above due to the expanded vocabulary. Similar gains are observed at the value 1.0, which corresponds to training the language model on the acoustic training set alone. Clearly some benefit is obtained in each case through a merging of the vocabularies between domains. Further gains can be found at intermediate values of the interpolation constant with a peak at about $\lambda = 0.75$. These results validate the statistical filtering used to select the CNCS collection. We find improvement both from an enlarged vocabulary and from

TABLE IV
CZECH ASR PERFORMANCE WITH LANGUAGE MODEL DOMAIN

LM	# types	OOV	WER
LM-Tr-C	43.9k	0.0%	40.23 %
LM-Tr	43.7k	5.8%	45.91 %
LM-LN	60.7k	8.8%	59.75 %
LM-CNCS	61.0k	6.2%	52.99 %

increased predictive power obtained by merging the filtered language model with an entirely in-domain language model. Table V shows the gains that can be obtained with unsupervised MLLR using two regression classes, for speech and for silence.

F. Analysis of Factors Affecting MALACH ASR Performance

We have presented in the Section I an example showing the nature of the highly spontaneous speech found in the interviews. The ASR results presented in Sections II-E.1 and II-E.2 illustrate the difficulties encountered due to OOVs, multiple languages, and heavily accented, elderly speech. As discussed in Section II-B, the noisy environments combined with poor articulation pose difficulties even for human transcribers. We provide here further analysis of the influence of these factors on ASR performance.

1) Speaker Population: Many of the survivors were originally from regions where the language in which they gave their interview (for example, English or Russian) was not frequently spoken. This resulted in heavily accented speech. It also resulted in speaking styles (choice of words) that were influenced by their place of origin. The problem is compounded by the fact that many of the interviewees traveled extensively during the course of their lives and in the process learned many languages; most of the interviewees claim proficiency in four to five languages.

A surprising observation follows from segregating the results of the third row in Table III into performance over interviewers and interviewees: the WERs for the interviewee speech in English is 39.4% while the rate for the interviewer is 47.9%. This is almost certainly due to the limited amount of interviewer speech within the interviews (typically less than 20%; 9% on average). This performance difference strongly suggests that the interviewers and interviewees are in fact from very different populations. This may pose a serious problem for cataloging applications since many spoken archives are recorded in the form of interviews, and the questions posed by the interviewers are at least as important as the responses by interviewees.

2) Spontaneous and Colloquial Nature of the Interviews: The interviews were conducted with the specific purpose of covering certain basic topics that would provide an insight into human experiences during a period of time in history. To this extent, there is a structure to the interviews, beginning with biographical questions followed by questions on education, occupation, living conditions, life in the camps, liberation, etc.. Therefore the interview segments have some initial awkwardness that gives way to spontaneous speech filled with disfluencies and sometimes whispered speech combined with nonspeech events such as crying and laughter, as they get into the middle of their stories. There are sections of frequent interruptions by the interviewer, sometimes to assist the interviewee along, and these rapid speaker changes and cross talk pose problems for recognizers. Background speech and frequent interruptions pose problems for adaptation to the interviewee's speaking style.

TABLE V
CZECH ASR PERFORMANCE FOR THE MERGED LANGUAGE MODEL LM-Tr-CNCS, MERGED VOCABULARY (81.9 k WORDS). UNSUPERVISED MLLR RESULTS ARE FOUND AT $\lambda = 0.75$

λ	WER
0.00	51.10%
0.25	46.68%
0.50	44.31%
0.75	43.92%
1.00	44.99%
MLLR	39.40%

The following experiment illustrates the extent of the spontaneous nature of the speech and the difficulties in modeling the speaker population. Two native English speakers were asked to read an hour of transcripts from the English interviews. An ASR system trained on BN data (without any VHF data in the acoustic or language models) achieved a WER of approximately 12% on the transcripts as read by these speakers (the baseline WER of this system on the F0 portion of the BN development test set was 18%). However, this system had a WER of *greater than 80%* on the same hour of speech from the VHF collection. This clearly illustrating the problems with the speaking styles and acoustics seen in this corpus.

The problem is compounded in Czech, where spontaneous speech is often colloquial in nature, and differs significantly from the standard Czech as used in writing or in broadcast media. This is further compounded by the usual stylistic differences seen between spoken and written material. The results in Tables II and IV clearly indicate the relatively large gains obtained for both English (4% absolute) and Czech (6% absolute) when merging the language model with data from other relevant domains.

3) Foreign Words: Many of the ASR errors are due to OOVs, a good number of these are foreign words, names, places and sequences of words uttered in a foreign language (such as German, Yiddish, or Hebrew). In English interviews, given that English is not the native language for almost all of these speakers, there is a tendency to switch to their native language whenever the interviewee cannot find the appropriate term in English, such as when describing cultural events. Therefore, detecting uncued switches between languages such as Yiddish, German, Polish, and English is essential. Many of these place names constitute cities, streets and names of concentration camps mentioned during the course of an interview. Although a lexicon can be built with the most frequently occurring words in this corpus, as the number of interviews processed grows, many new words will need to be added. It is very important for these words to be recognized correctly in order to aid subsequent retrieval technologies as explained in Section III.

G. Steps Toward the Automatic Transcription of the Entire VHF Archives

Processing of 500 h of English interviews has been completed and is available for NLP and search studies. A major challenge in decoding such volumes of data is the time taken for

actually doing so. The decoding process (including adaptation to speakers) consists of pre-processing steps that include the selection of the appropriate channel to be processed (VHF data is recorded in stereo), acoustic segmentation of the audio into speech and silence/noisy silence segments, clustering to generate coherent-speaker clusters, and efficient storage for processing. In order for this to be a viable solution, the entire process must be completed within 10xRT. Our current large vocabulary speech recognition system requires a segmentation of audio prior to recognition: this is a prerequisite for practical, rather than algorithmic reasons. Separation and removal of nonspeech segments will remove many insertion errors and provide robustness to the background noise during interviews. Also, the decoder benefits from segmentation in two ways: short segments reduce per-segment computational load for our current decoder implementation and eliminating nonspeech segments reduces the overall computational load. In addition, speaker labeling of segments allows adaptation to be performed on speaker-coherent clusters, which will further improve performance as illustrated in Section II-E1. Imperfect segmentation raises the challenge of automatically grouping (possibly speaker-impure) segments into clusters prior to adaptation; poor automatic groupings may impact gains from speaker adaptation. A detailed discussion of our progress toward identifying automatic schemes for the segmentation and speaker clustering process giving good end-result recognition performance is given in [26]. Detailed descriptions of the decoder used for processing these 1000 h are presented elsewhere [27], [28].

III. The NLP Components

In our NLP work, we have focused on three technologies named Entity Detection, Document Segmentation, and Segment Categorization.

Automatic Named Entity Detection plays two separate roles in the project. First, identifying named entities may assist the user by increasing the readability of the material presented by various search-oriented components. Second, named entity features can be applied to improve performance of other machine-learning components, as suggested by some of the results of our Topic Tracking work done in NIST's TDT evaluation [29].

The purpose of Document Segmentation is to partition the interviews into shorter (a few minutes long), topically coherent segments to be used as retrieval and annotation units. In contrast to previous Story Segmentation work described in [30], where the goal was to divide the transcript of news programming into individual stories, Document Segmentation deals with a problem of identifying topic boundaries *within* a transcript of a single interview. Unlike the blocks of BN programming, the structure of the interview is not based on a script containing predefined story boundaries, but allows spontaneous changes of the topic. This difference causes multiple issues influencing segmentation performance, such as less pronounced topical boundaries between the VHF interview segments than between news stories, less topical variety between the segments, and wider variance of length between the segments.

Automatic Segment Categorization aims to associate each interview segment with multiple semantic categories. The categories can be used directly by the end user or applied in high-level search components.

The VHF collection provides unique opportunities for research work in the three mentioned areas thanks to substantial amounts of data manually annotated with segment boundaries and semantic categories, and, as shown below, poses a real challenge to achieve a practical level of performance, owing to the richness and inherent heterogeneity of the data.

A. Named Entity Detection

We adopt an HMM model [31] to classify 31 categories of entities [32]. We tailored the system for VHF transcribed text, and annotated a 50 000 word development test corpus that comes from 30 random distinct speakers. We measure system performance by F_{3X}, the standard 3X (extent x content x type) F-measure, and slot error rate (SER) [33]. To enhance the usage of our text data, we apply domain-specific rule-based preprocessing to remove noise tokens and detect semantic sentence boundaries. This contributes 10% relative improvement to system performance.

We annotated 460 000 words of VHF transcription as training data for our system, to attain $F_{3X} = 72.4\%$ and $\mathrm{SER} = 33.7\%$. In contrast, a baseline system trained on annotated newswire text of 1 million words attains only $F_{3X} = 58.6\%$, $\mathrm{SER} = 48.3\%$. Furthermore, even a combined system that trains from both sources of data attains only $F_{3X} = 72.7\%$, $\mathrm{SER} = 35.6\%$, showing relatively small improvement in F_{3X}, yet large degradation in SER.

The results affirm the importance and value of high quality in-domain data in the corpus-driven Named Entity Detection approach. In fact, even the addition of large quantities of lower quality (out-of-domain) data do not necessarily help improve performance. Error inspection indicates that our named entity detection is affected by the rather limited quantity of training examples, especially in a domain of spontaneous and unpredictable speech, which hampers the system's generalization ability. We conclude that our system can further benefit from additional in-domain training data.

B. Segmentation

1) Statistical Models: For this study, we have extended our previous work done as part of the TDT project [30]. Our segmentation algorithm uses a combination of two probabilistic models, a DT and a maximum entropy (ME) model [34], to compute the probability of a segment boundary occurring at a given sentence boundary determined during the transcription process.

The DT model uses a combination of features, including presence of words and bigrams indicating segment boundaries (indicators are learned automatically from the training data by a mutual information criterion) and features comparing the distribution of nouns on the two sides of the proposed boundary. The ME model uses the features used by the DT and individual words, bigrams, and trigrams.

2) Training and Test Data: Our training data set consists of 15-min intervals extracted from 710 interviews, which represents 177.5 h of speech. The test set is based on four full interviews, representing 7.5 h of speech. Using timing information, we aligned the transcribed text with the segment

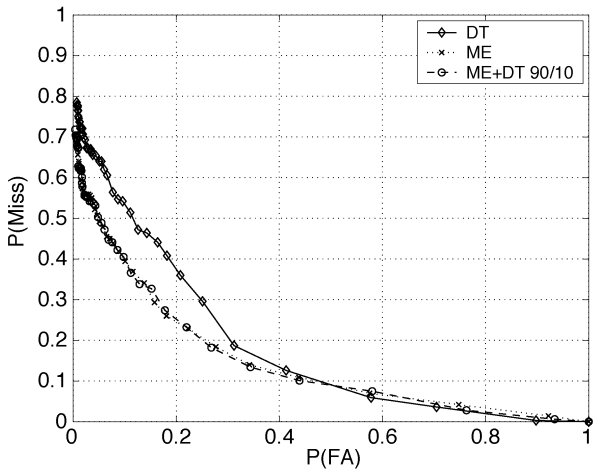


Fig. 3. Document segmentation: DT versus ME models.

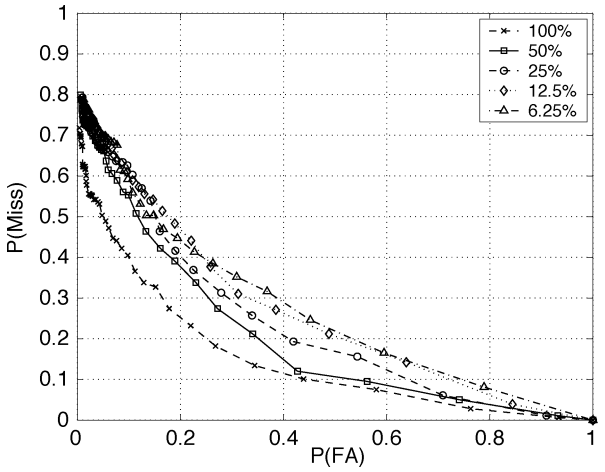


Fig. 4. Document segmentation: training data size.

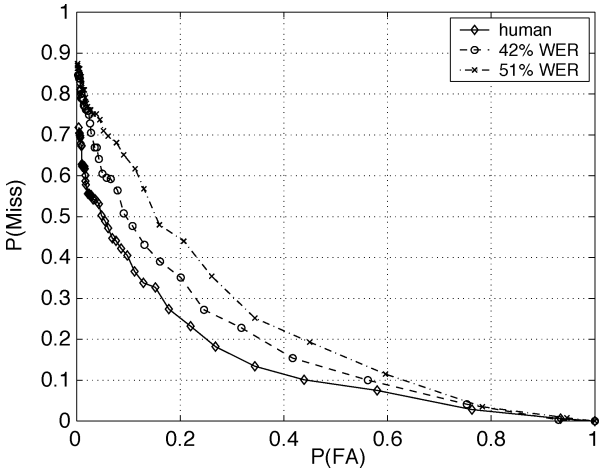


Fig. 5. Document segmentation: human versus ASR transcripts.

boundaries constructed by VHF, creating a training set of 2856 segments and a test set of 168 segments. We used sentence boundaries established by the transcription process (Section I) in experiments with both human and ASR transcripts. Our previous results [43] show only a small performance degradation when using silence-based sentence boundaries on ASR transcripts.

3) Measuring Segmentation Performance: To measure segmentation performance we used an approach similar to the one applied in the TDT segmentation task [35]. The performance measure is based on determining the agreement between computed and reference boundaries, using a sliding window moved through the segmented data. At each position of the window we declare correct segmentation if there is both a computed and a reference boundary or neither a computed nor a reference boundary found in the window. Similarly, a *false alarm* is declared if there is a computed boundary and no reference boundary in the window, or a *miss* is declared if there is no computed boundary and a reference boundary in the window. In contrast to the technique used in the TDT project, where the above described computation is performed for every word position in the data, we move the window so that the computation is done only with windows centered at sentence boundaries. The window length is set to ten words.

4) Comparing Segmentation Performance of the DT and ME Models: Fig. 3 compares segmentation performance of the DT and ME models. We observe that the ME model outperforms the DT model over most of the range of operating points. We observed a similar effect when segmenting BN stories [30]. Fig. 3 also shows results based on combining the scores of the DT and ME models in a linear way, putting 90% of the weight on the ME model. Using a combined model brought modest, but consistent performance gain in segmenting BN stories, whereas there is no improvement on topic segmentation of the VHF data.

5) Effect of Training Data Size on Segmentation Performance: To measure how segmentation performance depends on the size of the training set, we subsampled the available training data in a pseudo-random fashion to obtain smaller sets. Fig. 4 shows that the performance keeps improving as the training data size grows and that the size of the available set does not approach a point where the benefit of larger training set starts to diminish.

6) Segmenting Human and ASR Transcripts: Fig. 5 compares segmentation performance on manually and automatically transcribed data. The relative degradation caused by speech recognition errors is quite uniform over a wide range of operating points. Table VI compares performance degradation caused by ASR errors at the segmentation equal error rate (EER) operating point. ASR transcripts with 42% WER experienced less than half the performance degradation observed with 51% WER.

C. Categorization

1) Ground Truth: An important part of the cataloging process carried on by the Visual History Foundation was the assigning multiple category labels to segments of interviews. VHF uses a set of over 32 000 categories, ranging widely from easy-to-define entities as geographical names to broader concepts, e.g. various political or psychological characteristics. Examples of category names can be found in the first column of Table VII. The categories are organized into a hierarchy with some cases of multiple inheritance. A single interview segment

is frequently assigned to multiple categories, the average number of categories per segment in the data we experimented with is 5.5.

2) K-Nearest Neighbors Categorization: We selected a well-studied [36] k-nearest neighbors (or kNN) technique in our initial categorization experiments. The kNN approach is based on the premise that a test segment is likely to belong to the same categories as the training segments similar to it. Given a test segment, the algorithm first finds its k closest training segments (k nearest neighbors). To compute similarities between segments we used a symmetrized version of the Okapi formula [37], with morphological stem unigram features applied to represent the individual segments. The categories associated with the k segments are assigned scores, equal to summations of similarities between the individual segments and the test segment. Finally, the list of candidate categories is sorted by the scores and the list is thresholded to select the category assignments of the test segment.

3) Categorization Using ME Model: In ME models for segment categorization, the individual word n-gram features are used to predict the likelihood of a segment being assigned to a category. We experimented with two versions of the model: the first one is set up as a multi-categorizer and computes the likelihood of the current segment with respect to all the categories, the second one consists of separate binary models for the individual categories. Table VII shows examples of categories together with feature words selected by the training process to represent them.

4) Categorization Experiments and Results: In our categorization experiment we used a training set based on 683 15-min intervals of English interviews for which we have human transcripts available. This data represents 2618 categorized segments (1.3 million words). The test data consists of 334 segments (129 000 words).

We measure categorization performance using the common precision/recall approach. To obtain a single performance indicator, we compute F_1, the harmonic mean of precision and recall. We compute both micro-averaged scores, where all the individual category assignments contribute equally to the overall score, and macro-averaged scores, where the precision, recall and F_1 measures are first computed individually for each category and then averaged over the categories, making equal the contribution of the individual categories. To compute macro-averaged performance indicators based on a set of categories with varying number of corresponding correct segment assignments, individual category precision values are interpolated to a set of standard recall levels [38]. We also compute segment-based macro-averaged scores, in which the precision and recall values are computed separately for each segment and then averaged and interpolated in the similar fashion as the category-based macro-averaged scores.

Based on our initial categorization experiments investigating the effect of number of training segments on categorization performance (Fig. 6), and the fact that currently available training set is roughly eight times larger than the test set, we include in our performance computation only the categories with at least ten training samples available. Out of 323 such categories, only 247 are represented in the test set, the remaining 76 nevertheless contribute to the overall scores by potentially reducing the precision in the micro-averaged computation; their precision is set to zero for the whole range of recall values in the macro-averaged computation.

TABLE VI
SEGMENTATION: HUMAN AND ASR TRANSCRIPTS

Transcripts	EER[%]	Relative degradation[%]
Human	23	-
42% WER	26	14
51% WER	30	33

TABLE VII
ME MODEL FEATURES

category	training segments	feature words
family life	131	love, born, grandparent, birth, maiden, play
living under false identity	129	hide, paper, priest, Catholic, Warsaw
forced labor in the camps	96	work, camp, barrack
convents and monasteries	10	convent

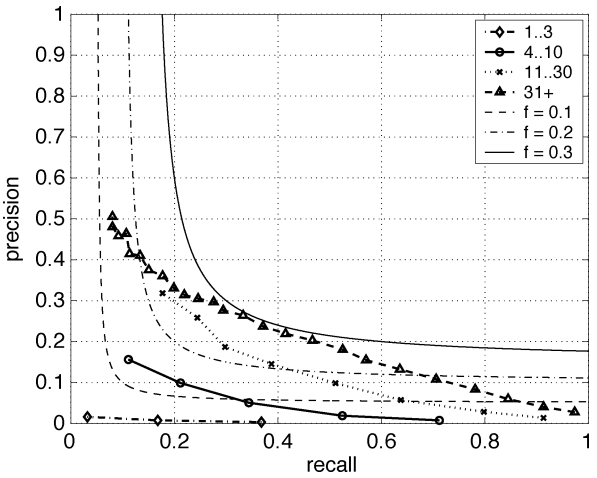


Fig. 6. Categorization: number of training segments per category, kNN, human transcripts, micro-averaged.

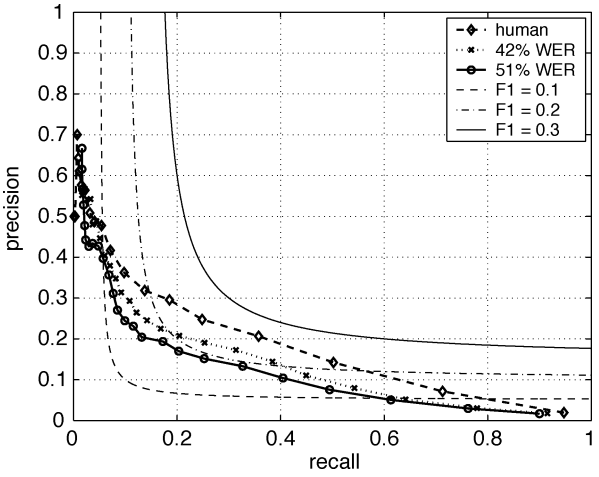


Fig. 7. Categorization: kNN, micro-averaged.

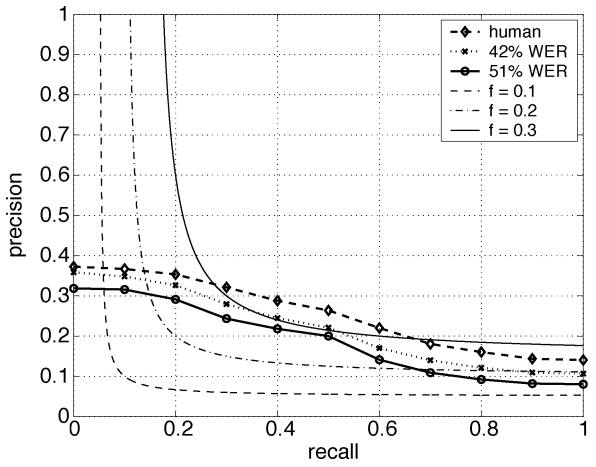


Fig. 8. Categorization: kNN, macro-averaged categories.

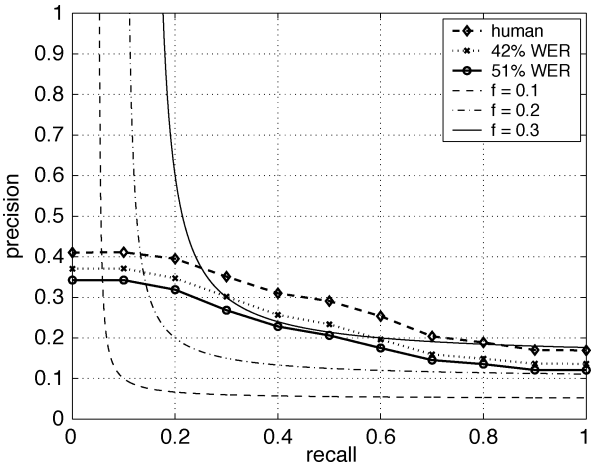


Fig. 9. Categorization: kNN, macro-averaged segments.

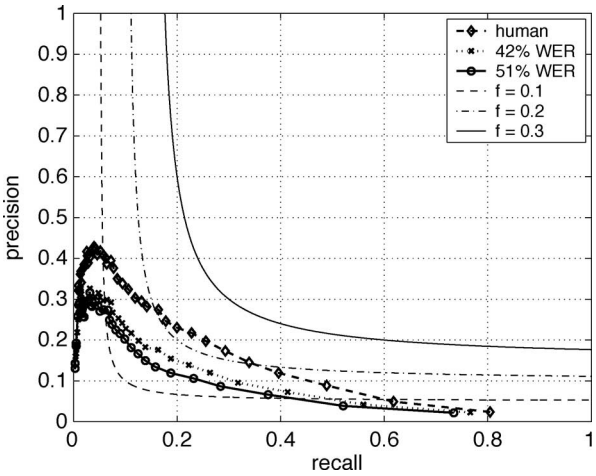


Fig. 10. Categorization: ME, micro-averaged.

Figs. 7–12 and Table VIII summarize the categorization results and illustrate the following trends.

- Current performance level on the given category set does not reach the point allowing broad practical application.
- KNN-based categorization outperforms the ME model under all the tested conditions.

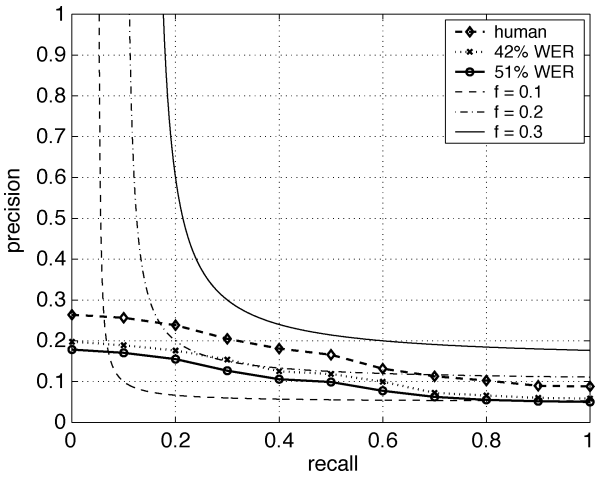


Fig. 11. Categorization: ME, macro-averaged categories.

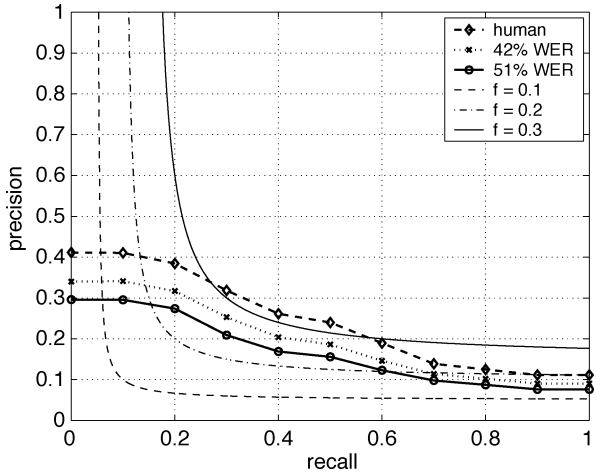


Fig. 12. Categorization: ME, macro-averaged segments.

- Speech recognition errors cause substantial performance degradation. ASR transcripts with 42% WER suffer more than half the performance degradation observed with 51% WER.

Figs. 13 and 14 contrast the performance of a single ME model trained to cover multiple categories, and a set of ME models trained to make binary decisions about the individual categories. Measured by the micro-averaged scores, dominated by the frequent categories, the single model outperforms the set of binary classifiers, whereas measured by the category-based macro-averaged scores, which assign equal weight to all categories, the set of binary classifiers outperforms the single model.

5) Analyzing Categorization Results: To analyze the categorization results, we examined the categorization algorithms at operating points yielding the best F value for the individual categories and segments. We were not able to identify category or segment properties linked in systematic and quantifiable way to performance level. The difficult to handle categories are often the ones describing broader concepts as "*evasion*", "*flight preparations*" and "*liberation*", and categories specifying a location together with a time interval, such as "*France, May 10 1940–February 28 1942*". The system performed better on more

TABLE VIII
CATEGORIZATION: kNN, HUMAN AND ASR TRANSCRIPTS

	Micro-averaged		Macro-averaged Categories	
WER[%]	F_1	degradation[%]	F_1	degradation[%]
Human	0.262	–	0.345	–
42	0.223	15	0.306	11
51	0.189	28	0.286	17

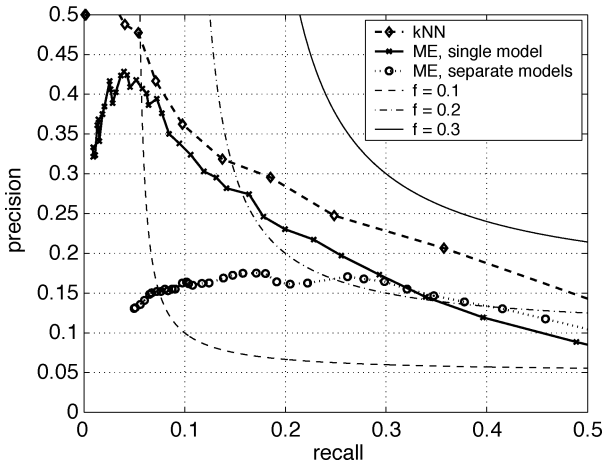


Fig. 13. Categorization: kNN, ME—single model, ME—separate models, micro-averaged.

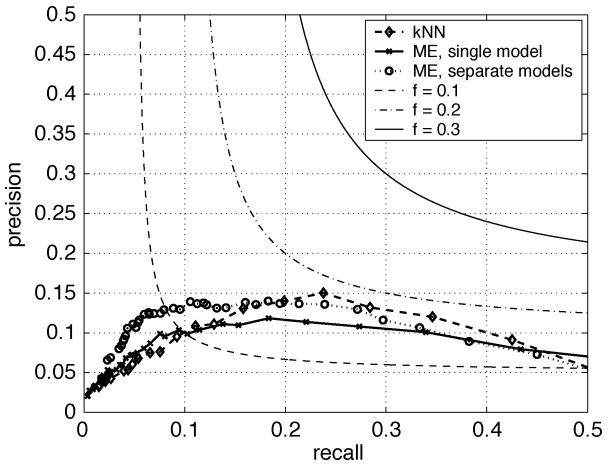


Fig. 14. Categorization: kNN, ME—single model, ME—separate models, macro-averaged categories.

specific categories, such as "*migration from Poland*" and categories describing a single location. Some of the poorly performing segments are very short.

IV. SEARCH ARCHITECTURE

The four capabilities described above, speech recognition, named entity recognition, topic-based segmentation, and topic classification, provide a sufficient basis for the design of interactive search systems [39]. In this section, we briefly highlight what is known about the needs of real users and then explain how these capabilities can be used together to meet those needs. We are currently implementing various components of a complete system.

A. User Needs

To assess user needs we conducted three sets of studies: 1) analysis of 600 written access requests received by VHF; 2) a week-long observation of the searching behavior of seven scholars from six disciplines in the humanities and the social sciences; and 3) a week-long observation of the searching behavior of nine experienced secondary school teachers. The most striking observation was the wide variety of users and uses. The mission of VHF is tolerance education, so it is not surprising that educators are well represented among the groups that we studied. The presence of historians and film producers is also not surprising. But there are also users with interests in anthropology, material and nonmaterial culture, linguistics, psychology (e.g., trauma studies), human rights advocacy, and law enforcement, among others. Two-thirds of the requests specified named entities (places, persons, and/or events); the remainder asked about more abstract concepts. Examples include *The liberation of Buchenwald and Dachau concentration camps*, *Sarah Ehrenhalt-Israel's life*, and *Motivational strengths that sustained survivors through the Holocaust*. Some searches required access to concepts that were not present in the thesaurus but that might be determined using speech recognition (e.g. find interviewees with fathers who were butchers). Some others involved concepts that might be difficult to assess using any technique (e.g., show only age-appropriate materials that would be suitable for classroom use).

We found that historians often required access to entire interviews, while other searchers (e.g., educators and film producers) needed specific passages. This points to a need for automatic segmentation in some cases. Moreover, we observed that users sometimes found it challenging to understand the context of retrieved segments, in part because stories are often told in an order different from the order in which events were experienced. This suggests that extraction of time expressions to support automatic generation of event timelines would be useful. We found that interviewer questions can be good markers when navigating an interview, that they provide useful access points for search, and that they can serve as useful aids for interpretation of what was said. This suggests that speaker change detection and speaker tracking would be useful capabilities.

Users were often observed taking copious notes, sometimes manually transcribing selected passages while watching interviews. This makes sense, since the information obtained from interviews will often be used to create of a written product (e.g., a book). If captured and shared in appropriate ways, these notes might be leveraged for the benefit of other users, or perhaps even as training data to improve the accuracy of automatic system components. This illustrates the benefits of transcriptions while placing the requirement for automated transcripts to be at a level of accuracy that will render them useful.

B. Search System Architecture

Fig. 15 shows the key components in a proposed interactive search system architecture. The three components on the left side of that figure have been described above; in this section we focus on the four components to the right. The index is a content-addressable store, typically implemented as a hash table. When

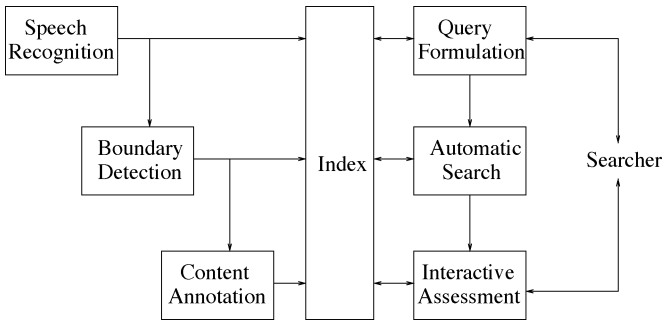


Fig. 15. Search system architecture.

presented with desired content (e.g., spoken words, normalized forms of person or location names, or automatically-assigned topic labels) the index returns a list of segments that contain or have been annotated with the specified content and a weight that represents the degree to which the specified content is describes the segment.

We view search as an iterative process performed by a searcher who uses the machine as a tool, rather than as a task performed autonomously by the machine. Robust facilities for query formulation, recognition of useful segments, and iterative query reformulation are therefore needed. Because many types of content features are indexed, functions specialized to each are needed in the query formulation component. Users can find thesaurus topic labels in two ways: by search, and by browsing. Approximate substring matching is a useful search technique when seeking a topic based on the characters contained in a label; free-text ranked retrieval techniques are more appropriate when searching based on definitions or scope notes (both of which are contained in the VHF thesaurus). If the search results in a conceptually related topic label, hierarchical browsing offers an alternative way of reaching the desired label. Similar functions can be provided for geographic locations, and map visualizations offer an additional way to specify desired locations. Similarly, timelines with range sliders offer a natural way of specifying a desired range of dates.

Person names pose unique challenges for systems based on speech recognition because in some cases they may not be present in the recognizer's vocabulary. We can minimize that effect for the VHF collection by including names found in the pre-interview questionnaires, but the prevalence of multiple transliterations for the same name and the presence of unanticipated names in some interviews result in a need for name searching based on phonetic similarity.

Searching based on spoken words, the third major type of indexed content, is somewhat more straightforward. The usual approach is to represent terms in a normalized form (e.g., automatically removing common endings to produce a stem) that matches the form that is indexed. In BN collections, it has proven to be useful to enrich the index with additional terms that exhibit similar usage in a collection of electronic text on the same subject [6], and it seems likely that similar techniques would be useful for spontaneous conversational speech as well.

When a fully formulated query is available, the automatic search component can then identify potentially useful documents. The search component must balance three desirable characteristics: high precision, high recall, and explainable behavior. These requirements are typically in tension; techniques that increase recall often do so at the expense of precision, and techniques such as blind relevance feedback that tend to limit this adverse effect often do so at the expense of explainability. Balancing the contribution from each type of indexed feature on the retrieval process presents a particular challenge; learned feature weights are practical in fully automated applications, but interactive applications can benefit from providing a greater degree of user control. Finally, three types of "proximity" features must be considered; temporal proximity in the interview (terms from adjacent segments can offer additional evidence for topicality), conceptual proximity in the thesaurus (is-a and part-whole relationships can serve as a basis for query expansion), and phonetic similarity (when searching for spoken words that are outside the recognizer's vocabulary).

Ultimately, searchers must judge which of the retrieved segments (or the interviews that contain those segments) meet their needs. In text, this is typically achieved by coupling a ranked list of brief surrogates with rapid access to the full text of each document. Unfortunately, replay of recorded speech is not rapid; with 3-min segments, listening to just the top ten retrieved segments would take half an hour. One way to overcome this limitation is to display two-level surrogates [40]; a brief surrogate for display in a ranked list, and a more extended surrogate that can be examined before committing to the replay of a selected segment. In our application, we can construct these surrogates using the recognized words, the identified named entities, the assigned thesaurus labels, and the pre-interview questionnaire; similar features have been found to be a useful basis for selection in previous studies [41]. Map and timeline displays could also be helpful, since locations and times were observed in our user studies to be important bases for selection in some cases.

In a design space this large, effective techniques for formative evaluation are essential. While user studies offer important insights, repeatable and affordable techniques for evaluating the effectiveness of the automated search component are also needed. We have therefore created an initial ranked retrieval test collection with 28 topics and over 600 h of automatically transcribed speech from interviews that were not used to train the ASR system. The 28 topics were based on actual user needs expressed in the 600 written requests to VHF that we examined, and relevance judgments were made using a search-guided assessment methodology that has demonstrated excellent coverage of the relevant document set in the Topic Detection and Tracking evaluations. This test collection fills an important gap in the available set of evaluation resources, augmenting the existing ground truth annotations that are available for evaluation of automated segmentation, classification, and named entity recognition.

C. Expanding the Design Space

The segment-then-label strategy described above draws on techniques that were originally developed for searching BN [2]. Since news programs are typically created by stringing together a sequence of stories, a segment-then-label decomposition is well matched to both the nature of the collection and to a clearly

understood user need (finding stories). The situation is far less clear in spontaneous speech, however. The interviews in our collection do exhibit a segmentable structure at a very broad scale; typically the first 20% addresses pre-war experiences, the middle 60% wartime experiences, and the remaining 20% post-war experiences. But finer-grained decompositions are more troublesome; it is sometimes hard to say where one topic ends and another begins. This problem is exacerbated by the limitations of our present understanding of the implications of different decisions about segmentation for the searcher; some uses may require extended segments in order to provide enough context (e.g., presentations intended to provoke classroom discussion), others might need only a specific statement (e.g., illustrating a point that has already been stated in a documentary film). If our initial "one size fits all" strategy proves unsatisfactory, we plan to explore alternatives in which we associate labels with spans of time independently of the spans assigned to other labels. This label-then-segment approach would make it possible to then assemble segments based on the minimum span that could satisfy the searcher's request [42].

The second half of this project will increase its focus on issues related to end user search, and many of these ideas presented here are currently being implemented.

V. Conclusions

Our overall goal is to provide significantly easier access to large collections of spontaneous conversational speech. In this paper, we described the creation of what we believe to be the world's first large-scale collection of spontaneous speech with the characteristics needed to evaluate information access technologies. Our ASR results on this collection are now sufficiently accurate to support downstream processing in two languages, English and Czech. We have created reference implementations for three key capabilities, named entity recognition, topic-based segmentation, and segment categorization, and we have explained how these components can be used together as a basis for interactive search. Our understanding of the requirements that guide this design has been informed by the results of multiple user studies, and we have developed a test collection that reflects those needs that can be used to characterize ranked retrieval effectiveness.

The research reported above has yielded three key insights: 1) adequate accuracy can be achieved by state-of-the-art recognition techniques to support many components of downstream processing of spontaneous conversational speech; 2) topic boundaries can be recognized reliably in spontaneous conversational speech, and human performance on that task can be approximated reasonably well by automated techniques; and 3) search-guided relevance assessment is a practical way of building a ranked retrieval test collection for spontaneous conversational speech. We are not yet satisfied with the categorization accuracy that we have achieved, and our analysis of those results points to a need for a more nuanced approach in which events, broad concepts, locations, and times are handled separately. Together, these insights inform our conceptual design for an interactive search system that can exploit the capabilities we have constructed, and that design in turn sharpens our understanding of the needed capabilities from each component. This, we believe, is the key to mastering the challenge of building systems that provide effective access to large collections of spontaneous conversational speech.

Acknowledgment

The authors gratefully acknowledge the advice on test collection development from S. Strassel of the Linguistic Data Consortium. The authors would like to thank J. Huang and L. Mangu from IBM and V. Venkataramani from Johns Hopkins for their contributions to this work.

References

[1] DELOS/NSF. (2003) E.-U. W. on Spoken-Word Audio Collections. [Online] Available: http://www.dcs.shef.ac.uk/spandh/projects/swag/

[2] J. Allan, Ed., *Topic Detection and Tracking: Event-Based Information Organization*. Boston, MA: Kluwer, 2002.

[3] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1992, pp. 517–520.

[4] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC spoken document retrieval track: a success story," in *Proc. 8th Text Retrieval Conf. (TREC-8)*, [Online] Available: http://trec.nist.gov, Nov. 1999.

[5] Survivors of the Shoah Visual History Foundation. [Online] Available: http://www.vhf.org

[6] A. Singhal and F. Pereira, "Document expansion for speech retrieval," in *Proc. 22nd Int. Conf. Research and Development in Information Retrieval*, Aug. 1999, pp. 34–41.

[7] S. Dharanipragada, M. Franz, J. S. McCarley, K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Statistical methods for topic segmentation," in *Proc. 6th Int. Conf. Spoken Language Processing*, Beijing, China, 2000, pp. 516–519.

[8] J. S. McCarley and M. Franz, "Influence of speech recognition errors on topic detection," in *Proc. 23rd ACM SIGIR Conf. Information Retrieval*, Athens, Greece, 2000, pp. 342–344.

[9] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication—Special Issue on Speech Annotation and Corpus Tools*, vol. 33, no. 1–2, pp. 5–22, Jan. 2000.

[10] J. Psutka, P. Ircing, J. Psutka, V. Radova, W. Byrne, J. Hajič, S. Gustman, and B. Ramabhadran, "Automatic transcription of Czech language oral history in the MALACH project: Resources and initial experiments," in *Proc. Text, Speech, and Dialog Workshop*, Berlin/Heidelberg, Germany, 2002.

[11] B. Ramabhadran, J. Huang, and M. Picheny, "Toward automatic transcription of large spoken archives—English ASR for the MALACH project," in *Proc. ICASSP*, Hong Kong, 2003.

[12] J. Huang, V. Goel, R. Gopinath, B. Kingsbury, P. Olsen, and K. Visweswariah, "Large vocabulary conversational speech recognition with the Extended Maximum Likelihood Linear Transformation (EMLLT) model," in *Proc. ICSLP*, Denver, CO, 2002, pp. 2597–2600.

[13] L. R. Bahl, P. de Souza, P. S. Gopalakrishnan, D. Nahamoo, and M. Picheny, "Robust methods for using context dependent features and models in a continuous speech recognizer," in *Proc. ICASSP*, Geneva, Switzerland, 1994.

[14] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP*, Philadelphia, PA, 1996, pp. 1137–1140.

[15] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Tech. Rep., CUED/F-INFENG/TR291, 1997.

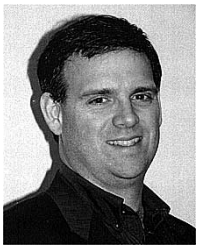
[16] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*: [Online] Available: http://htk.eng.cam.ac.uk/, 1997.

[17] P. Ircing, P. Krbec, J. Hajič, S. Khudanpur, F. Jelinek, J. Psutka, and W. Byrne, "On large vocabulary continuous speech recognition of highly inflectional language—Czech," in *Proc. 7th European Conf. Speech Communication and Technology (EUROSPEECH)*, Aalborg, Denmark, 2001.

[18] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Comput. Speech Lang.*, vol. 13, no. 4, pp. 359–393, 1999.
[19] J. Psutka, P. Ircing, J. V. Psutka, V. Radová, W. J. Byrne, J. Hajič, J. Mírovský, and S. Gustman, "Large vocabulary ASR for spontaneous Czech in the MALACH project," in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, 2003.
[20] R. Iyer, M. Ostendorf, and H. Gish, "Using out-of-domain data to improve in-domain language models," *IEEE Signal Processing Lett.*, vol. 4, pp. 221–223, Aug. 1997.
[21] W. M. Fisher, "Syllabification software," in *The Spoken Natural Language Processing Group, National Institute of Standards and Technology*, Gaithersburg, MD, 1976, [Online] Available: http://www.itl.nist.gov/div894/894.01/slp.htm.
[22] D. Kahn, "Syllable-based generalizations in english phonology," in *Indiana Univ. Linguistics Club*, Bloomington, IN, 1976.
[23] A. Sethy, B. Ramabhadran, and S. Narayanan, "Improvements in English ASR for the MALACH project using syllable-centric models," in *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU 03)*, Virgin Islands, 2003.
[24] M. Mohri, M. Riley, and F. C. Pereira, "Weighted finite-state transducers in speech recognition," in *Proc. Int. Workshop on Automatic Speech Recognition: Challenges for the Next Millenium*, Paris, France, Sept. 2000, pp. 97–106.
[25] A. Stolcke, "SRILM—An extensible language modeling toolkit," in *Proc. Int. Conf. Spoken Language Processing*, Denver, CO, 2002, pp. 901–904.
[26] B. Ramabhadran, J. Huang, U. Chaudhari, G. Iyengar, and H. J. Nock, "Impact of audio segmentation and segment clustering on automated transcription accuracy of large spoken archives," in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, 2003.
[27] G. Zweig, G. Saon, and F. Yvon, "Arc minimization in finite state decoding graphs with cross-word acoustic context," in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, 2003.
[28] G. Saon, G. Zweig, B. Kingsbury, L. Mangu, and U. Chaudhari, "An architecture for rapid decoding of large vocabulary conversational speech," in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, 2003.
[29] M. Franz, J. S. McCarley, T. Ward, and W.-J. Zhu, "Unsupervised and supervised clustering for topic tracking," in *Topic Detection and Tracking 2000 Workshop*, Gaithersburg, MD, [Online] Available: http://www.nist.gov/speech/tests/tdt2000/papers.htm 2000.
[30] S. Dharanipragada, M. Franz, J. S. McCarley, T. Ward, and W.-J. Zhu, "Segmentation and detection at IBM: Hybrid statistical models and two-tiered clustering," in *Topic Detection and Tracking: Event-Based Information Organization*. Norwell, MA: Kluwer, 2002.
[31] S. M. D. Bikel and R. Schwartz, "Nymble: a highperformance learning name-finder," in *Proc. Applied Natural Language Processing*, San Francisco, CA, 1997, pp. 194–201.
[32] A. Ittycheriah, M. Franz, and S. Roukos, "IBM's statistical question answering system—TREC-10," in *Proc. 10th Text Retrieval Conf. (TREC-10)*, Gaithersburg, MD, 2001, pp. 258–264.
[33] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *1998 DARPA Broadcast News Workshop*, Herndon, VA, 1998, pp. 249–252.
[34] A. L. Berger, V. D. Pietra, and S. D. Pietra, "A maximum entropy approach to natural language processing," *Computat. Ling.*, vol. 22, no. 1, pp. 39–71, 1996.
[35] J. G. Fiscus and G. R. Doddington, "Topic detection and tracking evaluation overview," in *Topic Detection and Tracking: Event-Based Information Organization*, 2002.
[36] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proc. 22nd ACM SIGIR Conf. Information Retrieval*, Berkeley, CA, 1999, pp. 42–49.
[37] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *Proc. 3rd Text Retrieval Conf. (TREC-3)*, Gaithersburg, MD, 1995, pp. 109–126.
[38] E. M. Voorhees and D. K. Harman, "TREC 2001 results," in *Proc. 10th Text Retrieval Conf. (TREC-10)*, Gaithersburg, MD, 2001, p. A-14.
[39] D. Soergel, D. Oard, S. Gustman, L. Fraser, J. Kim, J. Meyer, E. Proffen, and T. Sartori, "The many uses of digitized oral history collections: Implications for design," College of Inform. Studies, Univ. of Maryland, College Park, [Online] Available: http://www.clsp.jhu.edu/research/malach/pubs, 2002.
[40] J. Kim, D. Oard, and D. Soergel, "Searching large collections of recorded speech: A preliminary study," in *Proceedings of the ASIST Annual Meeting*. Medford, NJ: Information Today, 2003, pp. 330–339, to be published.
[41] A. Merlino and M. Maybury, "An empirical study of the optimal presentation of multimedia summaries of broadcast news," in *Automated Text Summarization*, I. Mani and M. Maybury, Eds., 1999.
[42] D. W. Oard and A. Leuski, "Searching recorded speech based on the temporal extent of topic labels," in *AAAI Spring Symp. Intelligent Multimedia Knowledge Management*, Palo Alto, CA, Mar. 2003.
[43] M. Franz, B. Ramabhadran, T. Ward, and M. Picheny, "Automated transcription and topic segmentation of large spoken archives," in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, 2003, pp. 953–956.

William Byrne (S'79–M'93) was born in New York, NY. He received the B.S. degree in electrical engineering from Cornell University, Ithaca, NY, in 1982 and the Ph.D. degree in electrical engineering from The University of Maryland, College Park, in 1993.

He is currently a Research Associate Professor in the Center for Language and Speech Processing and the Department of Electrical and Computer Engineering at The Johns Hopkins University, Baltimore, MD. His research is in the statistical processing of speech and language.



David Doermann received the M.Sc. degree in 1989 and the Ph.D. degree in 1993 from the Department of Computer Science, University of Maryland, College Park.

He is Co-Director of the Laboratory for Language and Media Processing, Institute for Advanced Computer Studies, University of Maryland, and an Adjunct Member of the Graduate Faculty. His research centers widely around the topics of document image analysis and multimedia information processing for digital libraries. He is active in the research community, has organized several symposia with government, industry, and academic participation, has served on numerous boards and program committees, and is an editor of the newly formed *International Journal of Document Analysis and Recognition*.

Martin Franz (SM'96) received the M.S. degree in computer science/electrical engineering from the Czech Technical University, Prague, Czech Republic, in 1986.

He joined the Continuous Speech Recognition Group, IBM T. J. Watson Research Center, Yorktown Heights, NY, in 1992 to work on language modeling for speech. Currently, he is with the Natural Language System Department at IBM Research, working on techniques for unstructured information processing.

Samuel Gustman received the B.S. degree in computer engineering from the University of Michigan, Ann Arbor.

He has been the Chief Technology Officer of the Survivors of the Shoah Visual History Foundation, Los Angeles, CA (http://www.vhf.org) since 1994. He also consults on video and digital library technology in Hollywood, CA.

Jan Hajič received the M.Sc. degree equivalent in computer science in 1984 and the Ph.D. degree in computational linguistics in 1994 from the Charles University, Prague, Czech Republic.

After working as an Research Assistant and later as an Assistant Professor at the Charles University since 1991, he became an Associate Professor and Head of the Institute of Formal and Applied Linguistics in 2003. He also worked as a Visiting Scientist at ISSCO, University of Geneva, Geneva, Switzerland, in 1990 and 1991, at IBM T.J.Watson Research Center, Yorktown Heights, NY, from 1991 to 1993, and as an Assistant Professor at the Johns Hopkins University, Baltimore, MD, from 1999 to 2000. His research interests include language modeling, morphology of inflective languages, machine translation, language understanding, and rich linguistic annotation.

Douglas Oard received the Ph.D. degree in electrical engineering from the University of Maryland, College Park.

He is an Associate Professor with a joint appointment in the College of Information Studies and the Institute for Advanced Computer Studies at the University of Maryland, College Park. His research interests center around the use of emerging technologies to support information seeking by end users, including speech retrieval, cross-language retrieval, and the exchange of implicit evidence of quality between networked users. He is an Associate Editor for *ACM Transactions on Information Systems*.

Michael Picheny (S'81–M'81–SM'97–F'01) received the Ph.D. degree from the Massachussetts Institute of Technology, Cambridge, MA.

He is the Manager of the Speech and Language Algorithms Group in the Human Language Technologies Group, IBM T.J. Watson Research Center, Yorktown Heights, NY. He has worked in the speech recognition area since 1981, joining IBM after finishing his doctorate. He has been heavily involved in the development of almost all of IBM's recognition systems, ranging from the world's first real-time large vocabulary discrete system through IBM's current ViaVoice product line. He has published numerous papers in both journals and conferences on almost all aspects of speech recognition. He is the co-holder of over 20 patents.

Dr. Picheny served as an Associate Editor of the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING from 1986–1989 and is currently the Chairman of the Speech Technical Committee of the IEEE Signal Processing Society. He has received several awards from IBM for his work, including three Outstanding Technical Achievement Awards and two Research Division Awards. He was named a Master Inventor by IBM in 1995 and again in 2000.

Josef Psutka received the M.S. degree in electrical engineering and the Ph.D. degree in cybernetics in 1974 and 1980, respectively, both from the Czech Technical University, Prague, Czech Republic.

He worked as an Assistant Professor at the Technical Institute,Pilsen, Czech Republic, from 1978 to 1991. In 1991, he joined the Department of Cybernetics, University of West Bohemia, Pilsen, as an Associate Professor, and became a Full Professor in 1997. His research interests include speech signal processing, acoustic modeling, large-vocabulary ASR, speech synthesis, and pattern recognition.

Bhuvana Ramabhadran received the Ph.D. degree in electrical engineering from the University of Houston, Houston, TX, in 1995.

She is a Research Staff Member in the Human Language Technologies Group, IBM T.J. Watson Research Center, Yorktown Heights, NY, and is currently the Co-Principal Investigator of MALACH from IBM. Since joining IBM in 1995, she has made significant contributions to the ViaVoice line of products. She has published in several journals and conferences and is the co-holder of several patents in the area of speech recognition. Her research interests include speech recognition algorithms, statistical signal processing, pattern recognition, and biomedical engineering.

Dagobert Soergel received the Ph.D. degree in political science from the University of Freiburg, Freiburg, Germany, in 1970.

He has been a Professor with the College of Information Studies, University of Maryland, College Park, since 1970. He has been working in the area of IR, specifically classification (taxonomy, ontologies) and thesauri, for over 40 years. He has authored two textbooks, regularly presents a tutorial on thesauri for knowledge-based assistance in searching digital libraries at JCDL and ECDL, and consults on the Harvard—Business Thesaurus.

Dr. Soergel received the American Society for Information Science Award of Merit in 1997.

Todd Ward is a Research Staff Member at the IBM T. J. Watson Research Center, Yorktown Heights, NY, where he manages a group focusing on multilingual natural language processing.

Wei-Jing Zhu was born in Guangzhou, China, and grew up in New York City. He received the B.A. degree in physics from Harvard College, Cambridge, MA, in 1991 and the Ph.D. degree in physics from Cornell University, Ithaca, NY, in 1998.

He has been conducting research in natural language understanding since 1999 at IBM T.J. Watson Research Center, Yorktown Heights, NY.