THE TUB 2006 SPOKEN TERM DETECTION SYSTEM

Jitendra Ajmera and Florian Metze

Deutsche Telekom Laboratories
Technische Universität Berlin
Berlin, Germany
{jitendra.ajmera|florian.metze}@telekom.de

ABSTRACT

This paper describes the Deutsche Telekom Laboratories' 2006 spoken term detection system submitted to the NIST 2006 Spoken Term Detection (STD) evaluation. The "indexing" system consists of a single unadapted pass of a speech recognizer on the test data, which outputs a confusion network. During "search", candidate positions for occurrences of individual search terms are determined using a confidence threshold. After adding terms not in the default dictionary using Festival, a new Viterbi alignment is then computed using the full term sequence at candidate positions and the relative durational behavior of different states of the HMM. This approach can theoretically find any expression containing at least one non-OOV word.

At evaluation time, the system reached an ATWV (Actual Term Weighted Value) of 0.32 on the merged Broadcast News, Conversational Telephony Speech, and Meeting development data.

1. INTRODUCTION

Deutsche Telekom Laboratories (T-Labs) [1], an "An-Institut der Technische Universität Berlin" (TUB) funded by Deutsche Telekom AG, recently started work on varied topics in the field of automatic speech processing using licensed software, but in-house system design and development.

In this paper, we present T-Labs first "keyword spotting" system, which was submitted to the NIST 2006 Spoken Term Detection evaluation as a stress test for the infrastructure.

A block diagram of our system is shown in Figure 1: the speech-to-text (STT) system is trained with JRTk [1]. The audio data is recognized using the Ibis decoder [2], converting recognizer lattices into confusion networks [3]. These are stored as an index for searching the spoken terms provided in the search term-list. The detector is based on posterior probability thresholding and the relative durational behavior of different states of the HMM.

[1] http://www.deutsche-telekom-laboratories.de/english/index.html

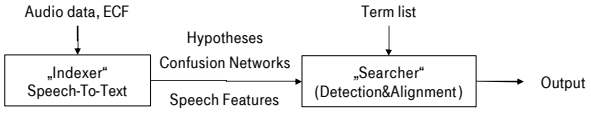

Fig. 1. Block Diagram of the TUB spoken term detection system.

This paper is organized as follows: Section 2 provides details of the training and other issues involved in different components of the speech recognizer ("indexer"), while issues related to search of spoken terms ("searcher") are covered in Section 3.

Section 4 gives a brief discussion of the system's performance on development and evaluation data using different performance indicators.

2. INDEXER TRAINING

The system was derived from the first pass setup of the ISL RT2004 "Meeting" system described in [4], using the same training setup.

2.1. Acoustic Model Training

Acoustic models for wide-band speech (16kHz) and telephony speech (8kHz) were trained on time-alignments generated on 16kHz close-talking training data (see Table 1) with simple models trained on Broadcast News data only. We then trained two systems using 6000 codebooks with merge-and-split Maximum Likelihood training on this time alignment and added two iterations of Viterbi training to estimate mixture weights for 24000 distributions. One of the systems used the 16kHz data directly, while for the other system the acoustic data was passed through a telephony filter and down-sampled to 8kHz. The phonetic decision tree (with context ± 2) was taken from the ISL RT2004 "Meet-

Corpus	Duration	Speakers
BN96+97	180h	3912
CMU	11h	93
ICSI	72h	455
NIST	13h	77
Total	276h	4537

Table 1. 16kHz acoustic training data: BN97 training data used automatically generated speaker clusters.

Beam	BN	CTS	CONFMTG	OVERALL
Word Error Rate				
Narrow	36.6%	59.0%	75.2%	57.2%
Wide	29.6%	56.7%	73.7%	53.7%
Lattice Nodes (in 1'000)				
Narrow	3'169	5'051	2'717	10'937
Wide	3'620	4'967	2'680	11'266
Items per Confusion Sets				
Narrow	1.48	1.58	1.63	1.56
Wide	1.49	1.62	1.62	1.58
Actual Term Weighted Value (AWTV)				
Narrow	0.28	0.18	0.14	0.26
Wide	0.38	0.21	0.13	0.32
Maximum Term Weighted Value (MWTV)				
Narrow	0.29	0.18	0.14	0.25
Wide	0.39	0.23	0.13	0.33

Table 2. Word Error Rate, ATWV performance with different beam sizes. Size of the index for different beam sizes.

ing" system. Each system used $\sim 300k$ Gaussians with diagonal covariance matrices.

Preprocessing consists of MFCC computation followed by the extraction of lMEL parameters, CMS/ CVN, stacking of ± 7 frames, LDA-based reduction of the dimensionality to 42 and a global STC transform. No VTLN was used, LDA and STC were computed on the 16kHz data only, but used in the 8kHz system, too.

Different beam configurations were tested as shown in Table 2. Word error rates were computed on a home-made reference derived from the RTTM file.

2.2. Language Model and Dictionary

Language models were also taken from the ISL RT2004 "Meeting" system [4]. This system used a 3-gram LM and a 5-gram LM with $\sim$800 automatically generated classes on a mixture of the Switchboard and Meeting transcriptions and also a 4-gram BN LM. All LMs were computed over a vocabulary of $\sim$47k words selected from BN, Meeting and SWB training data. Decoding and CNC uses a 3-fold context dependent interpolation of all three LMs.

2.3. Audio Segmentation

Original audio files were segmented into smaller speaker-specific segments in order to facilitate speech recognition. This was done using the modified Bayesian Information Criterion based approach proposed in [5]. In the development data, this segmentation resulted, on an average, in 10 second long segments. No silence detection, speaker normalization or channel adaptation was performed however.

3. SEARCHER

Confusion networks were obtained from word-lattices and provide a compact representation of alternative hypotheses. This framework is based on minimizing word-error-rate (WER) [3], which is better suited for spoken term detection task compared to MAP approach.

Confusion network, best hypothesis string and corresponding acoustic confidence scores were extracted for each utterance using the Ibis decoder and JRTk. Each word of a spoken term in the term-list was searched in the confusion network.

A confidence score for each spoken term is computed as geometric mean of confidence scores of individual sub-terms present in that spoken term. Confidence score of each individual sub-term, on the other hand, is computed based on presence of that sub-term in the best hypothesis. If a term is not found in the best hypothesis, its score is considered to be 1 minus score of the alternate hypothesis.

For example, an utterance corresponding to the spoken term *UNITED NATIONS* had the following confusion network and best hypothesis.

Confusion Network: ...THAT MANY OF THE { **NATIONS NATION'S** } ALL { ARE OUR } JUST...

Best-Hypothesis: ...{THAT(2) 599 617 0.831972} {MANY 618 657 0.999268} {OF(2) 658 669 0.993915} {THE 670 678 0.994} {**NATIONS 679 727 0.675**} {ALL 728 744 0.938725} {ARE 745 763 0.643445} {JUST 764 785 0.992460} ...

The confidence score for this term in the example (*UNITED NATIONS*) is then computed as $\sqrt{(1.0 - 0.994) \times 0.675} = 0.064$. A decision on existence of a spoken term is then made based on this confidence score, we used a threshold of 0.75.

An additional cue toward existence of a spoken term was also derived from analyzing the relative durational behavior of different states of the HMM that formed the spoken term. To achieve this, approximate begin and end-points for each spoken term were derived from the position of a sub-term in the confusion network and the best hypothesis. The search word sequence was then aligned in this window using Viterbi. Pronunciations for search terms missing in the

Beam	BN	CTS	CONFMTG
CN-based approach			
Narrow	0.28	0.18	0.14
Wide	0.38	0.21	0.13
"grep" approach			
Narrow	0.25	-0.03	-0.14
Wide	0.38	0.05	-0.09

Table 3. ATWV performance for a simple "grep" approach on the first best hypothesis and the approach based on confusion networks.

Set-up	BN	CTS	CONFMTG	OVERALL
Primary system				
Dev data	0.29	0.18	0.14	0.26
Eval data	0.39	0.16	0.05	N/A
Contrastive system				
Dev data	0.28	0.18	0.14	0.26
Eval data	0.39	0.16	0.03	N/A

Table 4. Performance in terms of ATWV for STD2006 development and evaluation data.

indexer dictionary were generated using Festival and added to the searcher before this alignment step. As we allow skipping of pauses in multi-word alignments and we assume the confidence of a word not present in the confusion set to be the "missing probability mass", this approach can in principle detect words that were not contained in the indexer's vocabulary. Using this cue with an experimentally optimized threshol results in an ATWV of 0.2853 on BN data for the narrow beam settings (instead of 0.2800).

The system exploiting this additional cue was submitted as the *primary* system, while our *contrastive* submission was based on acoustic confidence scores as explained above alone. Table 2 shows numbers on the development data for the contrastive system.

For diagnostic purposes, we also tested a simple "grep" approach on the first best hypothesis on the development data, the results are shown in Table 3.

4. EVALUATION SYSTEM

The evaluation data was processed with the settings optimized on the development data. Because of technical difficulties with out computer cluster, only the "narrow beam" setting could be submitted. As was the case for the development data, all data sets were processed using matching acoustic models, but otherwise identical systems and parameters. A comparison of our system's performance on development and evaluation data is shown in Table 4.

Other key performance indicators of our system on the evaluation data are shown in Table 5. Due to problems with the NAS server (Network Attached Storage) used in our computing grid, the timings reported are unreliable and significantly overestimating the actual compute time.

Indicator	Performance
Index size (MB/Hour of speech)	0.7730
Indexing time (Hours/Hour of speech)	4.3897
Search speed (Terms/Hour of speech)	0.1725

Table 5. System performance in terms of several indicators.

5. REFERENCES

[1] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal, "The Karlsruhe Verbmobil Speech Recognition Engine," in *Proc. ICASSP 97*. München; Germany: IEEE, Apr. 1997.

[2] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A One-pass Decoder based on Polymorphic Linguistic Context Assignment," in *Proc. ASRU 2001*. Madonna di Campiglio, Italy: IEEE, Dec. 2001.

[3] L. Mangu, E. Brill, and A. Stolcke, "Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks," *Computer, Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[4] F. Metze, C. Fügen, Y. Pan, and A. Waibel, "Automatically Transcribing Meetings Using Distant Microphones," in *Proc. ICASSP 2005*. Philadelphia, PA; USA: IEEE, Mar. 2005.

[5] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *IEEE Signal Processing Letters*, vol. 11, pp. 649–652, August 2004.