Technology and Corpora for Speech to Speech Translation
http://www.tc-star.org

Project no.:	FP6-506738
Project Acronym:	TC-STAR
Project Title:	Technology and Corpora for Speech to Speech Translation
Instrument:	Integrated Project
Thematic Priority:	IST

Deliverable no.: D6
Title: TC-STAR Recognition Baseline Results

Due date of the deliverable:	30th of September 2004
Actual submission date:	15th of October 2004
Start date of the project:	1st of April 2004
Duration:	36 months
Lead contractor for this deliverable:	IBM
Author:	Martin Westphal

Revision: [version 1.2]

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Project No: **IST-2002-2.3. 1.6**

Project: **Technology and Corpora for Speech to Speech Translation**

Acronym: **TC-STAR**

Document Title:	**TC-Star Recognition Baseline Results**
Deliverable Type:	**Report**
Version Number:	**version 1.2**
Status:	**Final**
Author(s):	**Martin Westphal**
Date:	**15 October 2004**
Security:	**Internal**
Keywords:	**Recognition, Baseline, Experiment, Evaluation**
Project Director:	**Gianni Lazzari**

Table of Contents

1 Summary

The following table gives an overview of baseline experiments performed by the project partners. It shows covered recognition tasks and languages and lists for each task the best word error rate (WER) that could be reached by one of the partner's existing recognition system.

Task	**Language**	**Participants**						**Best WER** segmentation:	
		IBM	IRST	LIMSI	RWTH	Sony	UKA	automatic	manual
LDC Broadcast News	English	✓	✓		✓	✓		17%	-
	Spanish	✓		✓	✓			19%	18%
	Mandarin			✓			✓	22%	-
TCStar_P	English	✓	✓	✓	✓		✓	40%	-
	Spanish	✓		✓	✓			44%	41%
EPPS	English	✓	✓	✓	✓		✓	-	32%
TED	English		✓				✓	31%	-

2 Introduction

This publication of baseline experiments in the work package 2 is meant to show the performance of available baseline systems for automatic speech recognition (ASR) that serve as a starting point for developing more advanced systems for the project.

The main focus is recognition accuracy. Usually only the first best sentence hypothesis generated by the speech recognizer is used as input to the translation component. So this first best sentence hypothesis will be evaluated in terms of word error rates.

Consecutive systems should be compared with the baseline numbers in order to document the progress. The baseline system of each partner used for this first evaluation should be kept and reused for any upcoming new test sets and conditions. This way the benefit of new or additional advanced technology during the lifetime of the project can be shown even if requirements of the evaluation might change.

2.1 Speech Recognition Metrics and Evaluation Procedure

The National Institute of Standards and Technology (NIST) has published tools and a methodology for measuring the accuracy of ASR systems. This approach has been successfully used over years for benchmarks and most partners of the project were already familiar with it. Therefore it was applied also for the baseline experiments reported in this document.

Each system was evaluated by measuring that system's word error rate (WER) except in Mandarin, where character error rate (CER) was the primary error measure. Word error rate is defined as the sum of the number of words in error divided by the number of words in the reference transcription. The words in error are of three types, namely ***substitution*** errors, ***deletion*** errors, and ***insertion*** errors. Identification of these errors

results from the process of mapping the words in the reference transcription onto the word in the system output transcription. This mapping was performed using NIST's SCLITE software package (http://www.nist.gov/speech/tools/index.htm).

- A substitution error results when the spellings of the reference word and the corresponding system output word differ.
- A deletion error results when the reference word has no corresponding system output word.
- An insertion error results when a system output word has no corresponding reference word.

The reference transcriptions are intended to be as accurate as possible, but there are necessarily some ambiguous cases and outright errors. The reference transcription for each turn was limited to a single sequence of words. This word sequence represents the transcriber's best judgment of what the speaker said. Segment time marks and corresponding reference transcriptions were provided in standard segment time marked (STM) format by the scoring sites.

The partner sites submitted their recognizer output as time marked conversation (CTM) files to the scoring sites. Before running the scoring scripts these files as well as the reference transcripts were normalized by applying a GLobal Mapping (GLM) file that was also provided by the scoring site. The results and the system description (provided by the participants) were sent to IBM in order to create this single report.

2.2 Languages, Tasks and Scoring Sites

The languages of interest for the baseline experiments are **English**, **Spanish** and **Mandarin**. The domains that are in focus of the project are **Broadcast News** (BN) and **Speeches**. Broadcast news data is usually provided in whole shows containing multiple speakers, different recording conditions and channels such as wide and narrow band. Also difficult background noise conditions can be found, e.g. music or babble noise. As automatic segmentation is not a key issue of the project, partners agreed to use either hand-labeled audio segments or the same automatic segmentation. Results might be given for a

Partitioned Evaluation (PE): using hand labeled, manually created segments

Unpartitioned Evaluation (UE): segments as well as speaker and/or channel labels are derived by an automatic process

Speeches are typically recordings containing only a single speaker with constant recording and channel conditions. Nevertheless speeches can be very difficult to recognize, especially in the task investigated here where non-native speakers occur.

There are 3 different tasks on which sites could submit recognizer output:

Task 1: existing publicly available data such as
- DARPA BN data (Hub 4) for English and Spanish
- TED data for English
- RT04 for Mandarin

Task 2: TCStar_P data,
10 hours for English and Spanish respectively,

sites that submit results using the TCStar_P data for adaptation should also submit results without adaptation

Task 3: European Parliamentary Plenary Speeches (EPPS), 1 hour of data collected during TCStar

Task 1 allows sites to show the accuracy of their systems for common, publicly available test data. Although this definition is very unspecific, there are only a few widely used benchmarks such as the broadcast news data sets that where used for previous evaluations by the Defense Advanced Research Projects Agency (DARPA) of the United States. Some partners had already systems available for this domain, others had to build them in order to have a starting point for the project. There is also training data available by the Linguistic Data Consortium (LDC) for this purpose. Training and test data for task 1 is US English, North American Spanish and Mandarin.

Task 2 consists of a subset of data collected during TC-Star_P that was defined by the corresponding scoring site. It turned out that there is a big mismatch between this data and the data used for training of the most systems. For Spanish, to give an example, the task 2 test data stems from Spanish radio stations whereas most of the available training data originates from American radio stations.

Task 3 consists of speeches from the European Parliament that were only available for English at this time.

The table below shows which task and language combination was scored by which site. Scoring sites supplied verified STM and GLM files and scored decoder output.

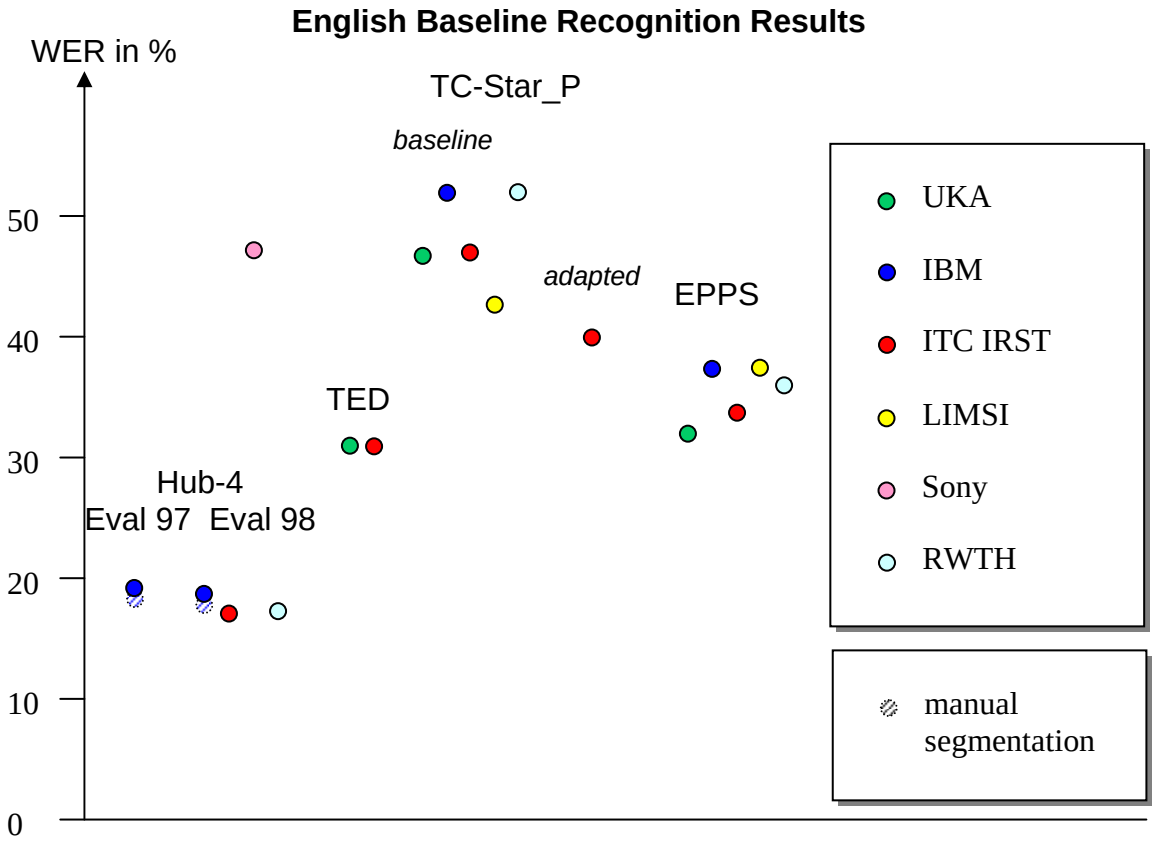
Task	English	Spanish	Mandarin
Task 1&2	IRST	LIMSI	UKA
Task 3	RWTH		-

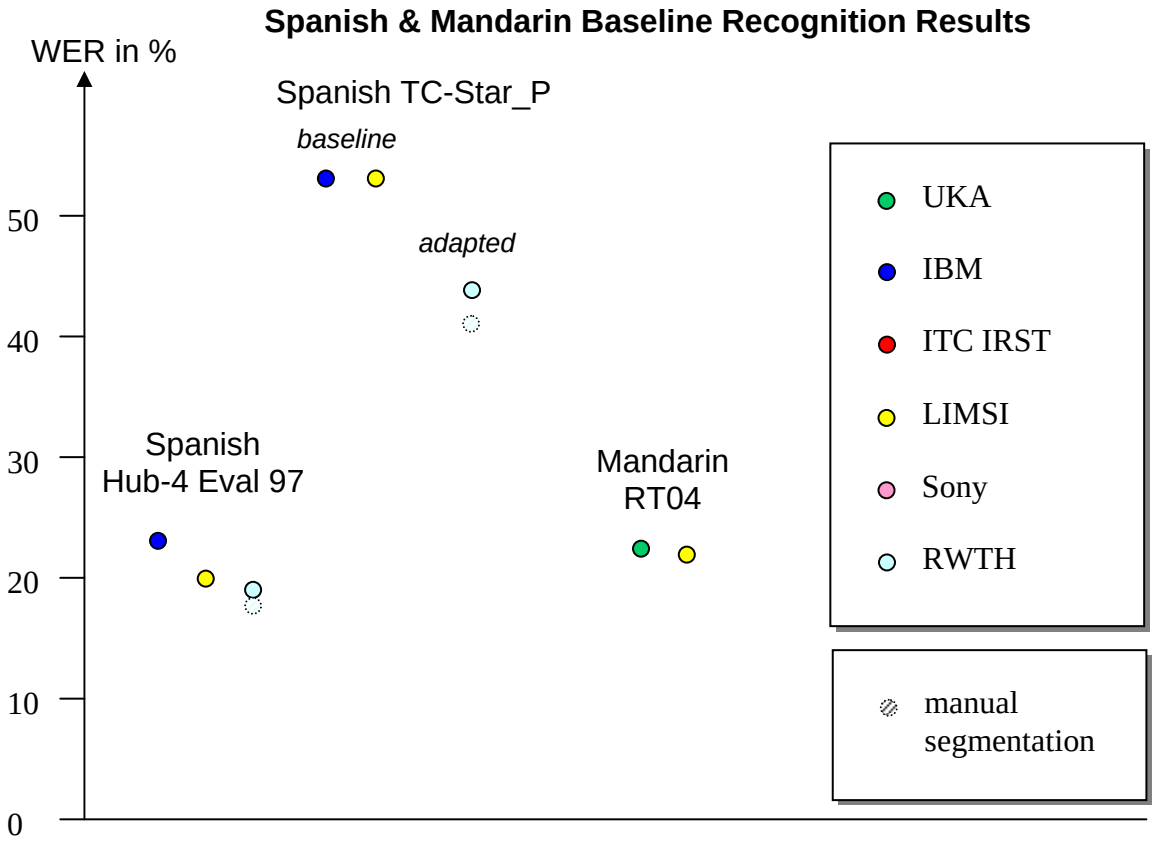
2.3 Participants and Task Coverage

The partners of work package 2 covered as many of the tasks as possible to provide a basis for future comparison. The table below gives an overview of the available recognition systems and results that can be found in the next chapters in more detail.

Partner	English Tasks	Spanish Tasks	Mandarin Tasks
IBM	1, 2, 3	1, 2	
ITC-IRST	1, 2, 3		
LIMSI	2, 3	1, 2	1
RWTH	1, 2, 3	1, 2	
Sony	1		
UKA	1, 2, 3	-	1

The following two figures give an overview as well, showing the error rates for most of the submitted recognition outputs (only some results for intermediate recognition steps were skipped). This gives a good impression about the challenge of each task and on the current status of the baseline systems.


English Baseline Recognition Results
WER in %
TC-Star_P
baseline
adapted
EPPS
TED
Hub-4
Eval 97 Eval 98
50
40
30
20
10
0
UKA
IBM
ITC IRST
LIMSI
Sony
RWTH
manual segmentation


Spanish & Mandarin Baseline Recognition Results
WER in %
Spanish TC-Star_P
baseline
adapted
Spanish
Hub-4 Eval 97
Mandarin
RT04
50
40
30
20
10
0
UKA
IBM
ITC IRST
LIMSI
Sony
RWTH
manual segmentation

3 Results English

3.1 English Results by Scoring Site

The following paragraphs give the scoring results as received by the scoring sites for the English recognition tasks. The scoring tool and the channel labels, given in the reference files, allowed reporting also for each so called F-condition. The given numbers are error rates for the total test set (Tot) or subsets with all the segments belonging to a certain F-condition such as prepared speech (F0), spontaneous speech (F1), low fidelity speech, including telephone channel speech (F2), speech in the presence of background music (F3), speech in the presence of background noise (F4), speech from non-native speakers (F5) and FX - all other speech.

Unless otherwise noted all numbers are word error rates (WER) in percent.

3.1.1 Results for Tasks 1 & 2 (scored and provided by ITC-IRST)

Scoring results for the English BN tasks and TED task

========== IBM ==========

TCSTAR_P

System	Tot	F0	F1	F2	F3	F4	F5	FX
UE_SI	60.0	40.4	45.6	59.3	38.3	65.4	73.5	83.9
UE_CMA	54.6	36.7	42.5	50.6	34.1	55.4	67.9	78.3
UE_MLLR	52.0	34.6	43.1	49.9	31.3	52.8	64.4	76.4

HUB4 Eval'97

System	Tot	F0	F1	F2	F3	F4	F5	FX
PE_SI	22.2	13.2	19.6	33.0	27.9	23.0	24.2	49.0
PE_CMA	19.2	11.4	16.8	28.8	24.4	19.8	22.6	41.6
PE_MLLR	18.2	10.9	16.0	27.5	24.3	19.3	21.1	37.2
UE_SI	22.4	13.5	20.7	32.3	28.3	24.8	24.1	45.7
UE_CMA	19.8	12.1	18.2	28.4	24.6	21.8	22.6	40.1
UE_MLLR	19.1	11.9	17.4	27.1	24.4	21.0	21.5	38.4

HUB4 Eval'98

System	Tot	F0	F1	F2	F3	F4	F5	FX
PE_SI	21.3	12.9	20.8	40.7	22.5	20.4	30.6	37.5
PE_CMA	18.5	11.2	18.4	37.3	20.0	16.6	29.8	33.3
PE_MLLR	17.8	11.0	18.3	35.1	19.9	16.1	29.8	30.6
UE_SI	21.8	14.0	22.2	32.3	24.2	20.9	28.1	37.0
UE_CMA	19.1	12.1	20.0	29.9	22.7	17.3	26.8	32.8
UE_MLLR	18.4	11.7	19.5	28.2	22.5	16.6	26.0	31.5

========== ITC-Irst ==========

HUB4 Eval 98

System	Tot	F0	F1	F2	F3	F4	F5	FX
B1	20.5	12.9	20.0	30.0	24.0	20.7	20.9	34.3
B2	18.7	11.7	18.7	24.7	23.0	19.2	19.6	30.7
S2	17.1	10.9	16.8	21.4	21.1	17.4	18.3	28.4

TCSTAR_P

System	Tot	F0	F1	F2	F3	F4	F5	FX
B1	54.0	39.6	44.3	46.1	35.9	59.5	64.0	74.5
B2	47.0	33.4	39.3	37.8	29.7	53.0	55.0	71.4
A2	40.1	26.1	33.0	36.4	23.7	47.4	46.9	65.0

TED

	Tot	cj29	dc57	fd29	hb64	ld29	ph50	ro31	yi59
B1	73.0	73.1	92.1	101.4	95.5	73.8	49.1	41.1	104.4
B2	62.2	61.9	76.2	95.3	81.7	60.6	37.6	33.7	99.0
A2	36.5	49.0	33.3	63.6	40.2	39.7	25.5	18.0	40.3
L2	31.2	33.4	32.8	57.7	38.8	35.7	24.9	15.2	28.8

========== LIMSI ==========

TCSTAR_P

System	Tot	F0	F1	F2	F3	F4	F5	FX
LIMSI	42.4	27.2	35.5	38.4	20.0	43.3	53.8	64.3

========== SONY ==========

HUB4 Eval 98

System	Tot	F0	F1	F2	F3	F4	F5	FX
SONY	47.4	41.0	42.9	57.5	48.4	49.8	68.9	59.5

========== RWTH ==========

HUB4 Eval 98

System	Tot	F0	F1	F2	F3	F4	F5	FX
RWTH	18.0	11.3	19.9	27.1	18.6	17.7	25.5	28.0
RWTH-MLLR	17.3	11.0	19.3	25.3	18.3	17.2	23.0	26.4

TCSTAR_P

System	Tot	F0	F1	F2	F3	F4	F5	FX
RWTH	52.0	37.3	33.0	47.9	31.7	53.9	62.9	74.3

========== UKA ==========

TCSTAR_P EN

System	Tot	F0	F1	F2	F3	F4	F5	FX
UKA	46.7	35.2	37.7	41.5	26.9	46.5	55.8	65.6

3.1.2 Results for Task 3: EPPS (scored and provided by RWTH)

Participant	baseline HUB-4 system WER[%]	after MLLR WER[%]	iterations
itc-irst	41.0	33.8	3
limsi	37.6	-	-
rwth	38.5	36.0	1
uka	32.0 *)	-	-
ibm	44.8/39.4 **)	37.5	?

*) Meeting Task Recognizer
**) after SI/CMA pass

3.2 English Results by Task

As it might be difficult to compare the numbers above against each other, given that the conditions as well as the system characteristics are very different, the numbers above are represented again below. They are now ordered by task and put together with system characteristics that were found in the system descriptions provided by the participants.

Only the best given results are listed in the tables below. The tag after the site name might indicate whether adaptation was used during decoding.
Training data used by most sites:
Audio Data:
Hub4 Hub4 1996 and Hub4 1997, 211 hours
- Hub4 1996 (LDC97S44)
- Hub4 1997 (LDC98S71)

Text Data:
Hub4-96 Hub4 1996 (LDC97T22)
Hub4-97 Hub4 1997 (LDC98T28)
Hub4-LM 1996 CSR Hub 4 Language Model corpus (LDC98T31), 130M words

3.2.1 Results for Hub-4 Eval 1997 (Task 1)

Site	**AM data**	**AM: tree size, prototypes**	**LM text**	**LM: 1-gram, 2-gram, 3-gram**	**vocab size (alternatives)**	**OOV rate**	**WER**
IBM PE	146h	8000, 128k (x4: band width dependent and SAT)	Hub4-LM, Hub4.96, hub4-97	103k, 1.9M, 2.3M	103k (109k)		18.2%
IBM UE	-"-	-"-	-"-	-"-	-"-		19.1%

3.2.2 Results for Hub-4 Eval 1998 (Task 1)

Site	**AM data**	**AM: tree size, prototypes**	**LM text**	**LM: 1-gram, 2-gram, 3-gram**	**vocab size (alternatives)**	**OOV rate**	**WER**
IBM PE	146h	8000, 128k (x4: band width dependent and SAT)	Hub4-LM, Hub4.96, hub4-97 = 133M words	103k, 1.9M, 2.3M	103k (109k)		17.8%
IBM UE	-"-	-"-	-"-	-"-	-"-		18.4%
ITC-IRST S2	143h	9000, 146k	LDC = 132M words		64k		17.1%
RWTH MLLR	96h	4000, 200k+350k			66k (78k)	1.3%	17.3%
Sony	104h +		Hub4-LM, WSJ, Web = 181M words	64k, 7.7M, 8.6M	24k (26k)		47.4%

3.2.3 Results for TED using an adapted system (Task 1)

Site	**AM data**	**AM: tree size, prototypes**	**LM text**	**LM: 1-gram, 2-gram, 3-gram**	**vocab size (alternatives)**	**OOV rate**	**WER**
ITC-IRST L2	143h + 8h	9000, 146k	mixed, adapted		64k		31.2%
UKA	300h	24k/6k, 300k	mixed = 237M words		25k	0.3%	31.0%

3.2.4 Results for TCStar_P using the baseline system (Task 2)

Site	**AM data**	**AM: tree size, prototypes**	**LM text**	**LM: 1-gram, 2-gram, 3-gram**	**vocab size (alternatives)**	**OOV rate**	**WER**
IBM UE	146h	8000, 128k (x4: band width dependent and SAT)	Hub4-LM, Hub4.96, hub4-97	103k, 1.9M, 2.3M	103k (109k)		52.0%
ITC-IRST B2	143h	9000, 146k	LDC = 132M words		64k		47.0%
LIMSI	150h	s1: 6300, s2: 12k, 188k s3: 12k, 375k (each x2: band width dependent)	Hub4+PSMedia, LDC newswire, LDC transcripts	-, 8M, 17M s3: 4-grams	65k (73k)		42.4%
RWTH	96h	4000, 200k+350k			66k (78k)	5.5%	52.0%
UKA	266h; 362h	24k/6k, 300k; 50k/10k,	SWB, Meeting, BN	3-gram, 5-gram	47k		46.7%

3.2.5 Results for TCStar_P using an adapted system (Task 2 optional)

Site	**AM data**	**AM: tree size, prototypes**	**LM text**	**LM: 1-gram, 2-gram, 3-gram**	**vocab size (alternatives)**	**OOV rate**	**WER**
ITC-IRST A2	143h + 6.5h	9000, 146k	LDC = 132M words		64k		40.1%

3.2.6 Results for EPPS (Task 3)

Site	**AM data**	**AM: tree size, prototypes**	**LM text**	**LM: 1-gram, 2-gram, 3-gram**	**vocab size (alternatives)**	**OOV rate**	**WER**
IBM MLLR	146h	8000, 128k (x4: band width dependent and SAT)	Hub4-LM, Hub4.96, hub4-97	103k, 1.9M, 2.3M	103k (109k)		37.5%
ITC-IRST MLLR	143h	9000, 146k	LDC = 132M words		64k		33.8%
LIMSI	150h	s1: 6300, s2: 12k, 188k s3: 12k, 375k (each x2: band width dependent)	Hub4+PSMedia, LDC newswire, LDC transcripts	-, 8M, 17M s3: 4-grams	65k (73k)		37.6%
RWTH MLLR	96h	4000, 200k+350k			66k (78k)	1.6%	36.0%
UKA	266h; 362h	24k/6k, 300k; 50k/10k,	SWB, Meeting, BN	3-gram, 5-gram	47k		32.0%

4 Results Spanish

4.1 Spanish Results scored and provided by LIMSI

```
           hub4   tcstar-p
IBM        23.3%  53.3%
LIMSI      20.0%  53.3%
RWTH       18.8%  43.9%
RWTH (c)   17.8%  41.3%

Notes:
- RWTH (c) is a contrastive system using manual segmentations of the signal
- RWTH's system training includes the tcstar-p data.  The others don't.
```

4.2 Spanish Results by Task

4.2.1 Results for Hub-4NE 1997 (Task 1)

Site	**AM data**	**AM: tree size, prototypes**	**LM text**	**LM: 1-gram, 2-gram, 3-gram**	**vocab size (alternatives)**	**OOV rate**	**WER**
IBM	20h	3000, 480k	SNT, Hub4, EP+HA	47k, 2.4M, 2.5M	47k	2.6% on eval	23.3%
LIMSI	30h	1600, (x4: gender and band width dependent)	LDC, Hub4, Caretas	-, 15M, 24M 4-gram rescoring	65k (79k)	1.4% on eval	20.0%
RWTH task1	30h	2500, 270k	SNT, Hub4		50k	2.1% on dev test	18.8%
RWTH task1 (manual segmentation)	30h	2500, 270k	SNT, Hub4		50k	2.1% on dev test	17.8%

LDC: 389M words, all newspaper and newswire texts by LDC
SNT: 140M words, newswire text by LDC (LDC95T9)
SNT-2: newswire text by LDC (LDC99T41)
Hub4: transcripts by LDC (LDC98T29)
EP+HA: articles from Spanish newspapers
Caretas: 9.6M words, online newspaper, recent

4.2.2 Results for TC-Star_P using the baseline system (Task 2)

Site	**AM training data**	**AM: tree size, prototypes**	**LM training text**	**LM: 1-gram, 2-gram, 3-gram**	**vocabulary (alternatives)**	**OOV rate**	**WER**
IBM	20h	3000, 48k	SNT, Hub4, EP+HA	47k, 2.4M, 2.5M	47k		53.3%
LIMSI	30h	1600, (x4: gender and band width dependent)	LDC, Hub4, Caretas	-, 15M, 24M 4-gram rescoring	65k (79k)		53.3%

4.2.3 Results for TC-Star_P using an adapted system (Task 2 optional)

Site	**AM training data**	**AM: tree size, prototypes**	**LM training text**	**LM: 1-gram, 2-gram, 3-gram**	**vocabulary (alternatives)**	**OOV rate**	**WER**
RWTH task2	30h + 7.5h	2500, 284k	SNT, Hub4, SNT-2, TCStar		12.5k full test set coverage	1.2% on dev test	43.9%
RWTH task2 (manual segmentation)	30h + 7.5h	2500, 284k	SNT, Hub4, SNT-2, TCStar		12.5k full test set coverage	1.2% on dev test	41.3%

5 Results Mandarin

Mandarin Results scored and provided by UKA:

For Mandarin we made use of the NIST RT04 test, including about one hour of broadcast news data from the following three sources: CCTV (20mn), RFA (20mn) and NTDTV (21mn). As is usually done by NIST for Mandarin, the recognition results are measured in terms of character error rate instead of word error rate.

	Character Error Rate
LIMSI:	22.0%
UKA:	22.4%

6 System Descriptions English

6.1 English BN Recognizer by IBM

IBM English TC-STAR Baseline System Description

1) PRIMARY TEST SYSTEM DESCRIPTION:

One single baseline system was run on the 3 evaluation tasks. For task 1, baseline numbers are provided for the DARPA BN Eval97 and Eval98 tasks, both with automatic data partitioning as well as supervised data partitioning. For task 2, baseline numbers are provided for the TC_STAR_P BN evaluation set, using only automatic partitioning. For task 3, baseline numbers are provided for the TC_STAR parliamentary data using supervised partitioning. In all cases involving automatic partitioning, the output from LIMSI's data partitioner[1] was used.

The BN system uses 60 dimensional feature vectors obtained from an LDA projection. The source space for the LDA projection is 117 dimensional and obtained by stacking 9 temporally consecutive 13 dimensional acoustic observation vectors. The vectors contain 12 cepstral parameters obtained from an inverse DCT of the log outputs of a 24 band, triangular filter bank. The filters are positioned at equidistant points on the Mel-frequency scale between 0 and 8 kHz. In addition to the 12 cepstral parameters, the vectors contain a raw frame energy parameter. The ceptral parameters are mean normalized on a per utterance basis. The energy parameters are translated based on the max energy, also on a per utterance basis.

The system uses 4 gender independent acoustic models. Each of these models are continuous density left-to-right HMMs using 16 component Gaussian mixture emission distributions and uniform transition probabilities. Each HMM has 3 states except for the silence HMM which is a single state model. The system uses 50 phones, 42 speech phones, 1 silence phone, 5 noise phones and 2 filled pause phones. The speech and filled pause HMMs use 7982 context dependent tied state distributions obtained by decision tree clustering of triphone statistics using context questions based on 77 phonetic classes. In addition, each model uses a global Semi-Tied Covariance (STC)[2,3] linear transformation.

First two 256-component text-independent Gaussian mixture models were built for wide (8kHz) and narrow (4kHz) band speech on the subset of the training that was labeled with this information. Then, the entire training set was classified as either wide or narrow band based on these mixture models. Then, the first acoustic model was built on all the data that was classified as wide-band data. A second narrow-band model was then obtained from the first model by MAP adaptation on the narrow-band data. The third model was obtained by Speaker Adaptive Training (SAT) using Constrained Model-space Adaptation (CMA)[2,3]. First, the wideband training data was partitioned such that each partition contained at least 10 seconds of data from a single speaker in a show. The data fragments that did not fall in any of the partitions were lumped together in a separate show-dependent cluster. Then a single CMA linear transform was

estimated for each cluster on all the speech frames in that cluster. After estimating the SAT transforms, the SAT model was obtained by single pass retraining, using the non-SAT model and LDA+STC transformed features for the expectation step and the SAT transformed features for the maximization step. The narrow-band model was obtained using the same process on the narrow-band data.

Decoding was performed in 3 passes and uses adaptation based on data clusters obtained either from the automatic partitioning or from the supervised speaker information. All passes use a single static decoding graph with 44M arcs and 20M states that was built from a trigram language model (103k 1-grams, 1.9M 2-grams and 2.3M 3-grams), a 103k word lexicon with 109k pronunciations and the HMM components. The acoustic models use cross-word contexts.

1) The bandwidth appropriate Speaker Independent (SI) model was used to get an initial transcript.
2) Based on the transcript from 1) and the SAT model, a CMA transform is estimated on a per cluster basis and the recognition process is repeated using the SAT model and those CMA transforms.
3) Based on the transcript from 2), the SAT model and the CMA transform, an MLLR[4] transform is estimated and the recognition process is repeated using the MLLR adapted SAT model and CMA transform.

2) ACOUSTIC TRAINING:

The acoustic models were trained on the Hub4 1996 (LDC97S44) and Hub4 1997 (LDC98S71) training sets, a total of 211 hours of recordings. Using the timing information in the corresponding training text corpora (LDC97T22 and LDC98T28), this data was processed to discard non-speech and overlapping speech segments and resulted in about 146 hours of usable speech. Furthermore, the transcripts were text normalized to about 1.6M words (with about 35k unique words). A training lexicon to cover the training set was derived from the Pronlex dictionary and manual augmentation.

3) GRAMMAR TRAINING:

The 3-gram language model is a Katz backoff model using Good-Turing discounting to reserve probability mass for unseen events and was built using the SRI LM toolkit[5]. An initial model was trained on the 1996 CSR Hub4 Language Model corpus (LDC98T31) (131M words after text normalization) and the text normalized acoustic training transcripts (1.6M words, included 4 times). The 103k lexicon was obtained by taking the 100k most frequent words in the LM training corpus and adding all words in the acoustic training text, not seen in the 100k vocabulary. The initial model, using that vocabulary, had 103k 1-grams, 7.2M 2-grams and 9.4M 3-grams and was subsequently shrunken using an entropy-based objective[6] to 103k 1-grams, 1.9M 2-grams and 2.3M 3-grams.

4) RECOGNITION LEXICON DESCRIPTION:

The 103k lexicon was obtained by taking the 100k most frequent words in the LM training corpus and adding all words in the acoustic training text, not seen in the 100k vocabulary. Pronunciations are based on a 50 phone set (42 speech, 1 silence phone, 5 noise and 2 filled pause phones). Pronunciations were obtained from the Pronlex lexicon and augmented with manual pronunciations.

5) EXECUTION TIME

Recognition experiments were run on a Pentium 4, 2.8GHz Xeon processor with 512kB cache and 2Gb memory. No particular attention was spend on optimizing the real time factor. Each pass runs at about 4xRt (12xRt overall for the 3 recognition passes).

6) REFERENCES

[1] J. L. Gauvain, L. Lamel and G. Adda, “The LIMSI Broadcast News Transcription System“, Speech Communication, 37(1-2), pp. 89-108, 2002.

[2] M. J. F. Gales, “Maximum Likelihood Linear Transformations for HMM-based Speech Recognition“, in Computer Speech and Language, No. 12, pp. 75-98, 1998.

[3] G. Saon, G. Zweig and M. Padmanabhan, “Linear Feature Space Transformations for Speaker Adaptation“, in Proceedings International Conference on Acoustics Speech and Signal Processing, Salt Lake City, UT, 2001.

[4] C. J. Leggetter and P. C. Woodland, “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs“, in Computer Speech and Language, No 9. , pp. 171-186, 1995.

[5] A. Stolcke, “SRILM - An Extensible Language Modeling Toolkit“, in Proceedings International Conference on Spoken Language Processing, Denver, CO, Sept. 2002.

[6] A. Stolcke, “Entropy-based pruning of backoff language models“, in Proceedings DARPA Broadcast News Transcription and Understanding Workshop, pp. 270-274, Lansdowne, VA, Feb. 1998, Morgan Kauffmann.

6.2 English BN Recognizer for Task 1a (Hub4) by ITC-IRST

ITC-irst speech recognition system

HUB4 BROADCAST NEWS EVAL-98 TEST

1) ACOUSTIC FRONT-END

All input speech data is segmented using a Bayesian Information Criterion. The maximum segment length is 50 seconds. Segments are classified into acoustic conditions. The result of this process is a set of speech segments with cluster, gender and telephone/wide-band labels. An automatic clustering is performed for all segments that belong to the same class. For clustering purposes, only automatic segmentation and classification is used both in training and recognition. During training, the manual transcriptions are aligned with cluster boundaries at the word level.

The acoustic features of the ITC-irst speech recognition system include 13 Mel-frequency Cepstral Coefficients (MFCCs) and their first and second order time derivatives into a 39-dimensional feature vector. The MFCCs are computed every 10ms using a Hamming window of 20ms length. The filter-bank contains 24 triangular overlapping filters which are centered at frequencies between 125 and 6750 Hz.

For the first and the second recognition pass two different acoustic front-ends are used. This is necessary as a supervised acoustic normalization technique is used in the second recognition pass (see section 2, below).

a) For the first pass, Cluster-based Cepstral Mean and Variance Normalization (CMVN) ensures that for each cluster the static features have mean zero and variance one.
b) For the second pass, Segment-based Cepstral Mean Normalization (CMN) is applied to the static features, adjusting the mean of each static coefficient for each segment to zero. No variance normalization is employed in this case.

2) ACOUSTIC MODEL:

We used the BN-E data released by the LDC in 1997 and 1998 for the training of the acoustic models. The corpora contain a total of about 143 hours of usable speech data.

The acoustic models are state-tied, cross-word, gender-independent, bandwidth-independent triphone HMMs. A phonetic decision tree is used for tying states and defining the context-dependent allophones. The system has about 9000 tied states and about 146000 Gaussians.

For training the acoustic models for the first recognition pass, a standard MLE acoustic training procedure is applied on the CMVN-transformed features.

For training the acoustic models for the second recognition pass, the acoustic normalization procedure described in [1,2] is employed:

a) a set of target models is trained on untransformed, mean-normalized feature vectors. The target models are tied-states triphone HMMs with a single Gaussian density for each state.
b) for each cluster in the training data, a CMLLR [3] transform is estimated w.r.t. the target models.
c) the CMLLR transforms are applied to the feature vectors. The resulting, transformed or normalized feature vectors are supposed to contain less speaker, channel, and environment variability.
d) a conventional ML training procedure is used to initialize and train the recognition models on the normalized features, including state tying and the definition of the context-dependent allophones.

3) LANGUAGE MODEL:

Trigram language models were trained on about 132 million words of broadcast news transcripts distributed by LDC and on the transcripts of the BN-E training data.

4) RECOGNITION LEXICON DESCRIPTION:

The pronunciations in the lexicon are based on a set of 45 phones. The lexicon contains 64k words. It has been generated by merging different source lexica (LIMSI '93, Cmudict, Pronlex). In addition, there is a model for silence and seven models for filler words and breath noises.

5) RECOGNITION PROCESS

In the first recognition pass the recognizer achieves a WER of 20.5%. The output of the first pass is used as a supervision for adaptation of the recognition models using the MLLR method [4]. For MLLR, two regression classes are used which have been determined in a data-driven manner. Mean vectors are adapted using full transformation matrices, while diagonal transformation matrices are used to adapt the variances. The reported results are the output of the second pass of the recognizer after three steps of adaptation. No normalization of the feature vectors except segment-based mean removal has been applied in the second pass, for details see [2].

A description of the cross-word decoding algorithm can be found in [5].

6) EXECUTION TIME:

The execution time of the first and second decoding pass is 255826.17 seconds (147485.82 seconds + 108340.35 seconds) on an Intel Xeon 2.4Ghz machine with 512KB cache and 4GB memory. This corresponds to 23.63xRT.

7) SUBMISSIONS SUMMARY:

For this task, we submit three results:

(B1) baseline: output of the first recognition step of the baseline system.

(B2) baseline+mllr: output of the second recognition step of the previous system, after three iterations of unsupervised MLLR adaptation. Supervision provided by the output of system (B1).

(S2) sup+mllr: output of the second recognition step using models trained with supervised acoustic normalization, after three iterations of cluster-based unsupervised MLLR adaptation, Supervision provided by the output of system (B1).

8) REFERENCES:

[1] D. Giuliani, M. Gerosa and F. Brugnara, “Improved Automatic Speech Recognition through Speaker Normalization“, to appear in ICSLP 2004.

[2] G. Stemmer, F. Brugnara, D. Giuliani, “Using Simple Target Models for Adaptive Training“, submitted to ICASSP 2005.

[3] M.J.F.Gales, “Maximum Likelihood Linear Transformations for HMM-based Speech Recognition“, Computer Speech & Language, Vol. 12, pp.75-98, 1998

[4] C.J. Leggetter, P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,“Computer Speech & Language, Vol. 9, pp. 171-185, 1995.

[5] F. Brugnara, “Context-Dependent Search in a Context-Independent Network“, ICASSP 2003, 360-363, Hong Kong, April 2003.

6.3 English BN Recognizer for Task 1b (TED) by ITC-IRST

ITC-irst speech recognition system

TRANSLANGUAGE ENGLISH DATABASE (TED) TEST CONDITION

1) ACOUSTIC FRONT-END

The acoustic front-end of the system that has been used for the TED corpus is the same as for the system employed for HUB4 BN EVAL-98 test condition, i.e. input speech is automatically segmented and clustered, the features are Mel-frequency Cepstral coefficients which have been normalized using Cluster-based Cepstral Mean and Variance Normalization (CMVN). Recognition is done lecture-by-lecture, i.e. each lecture is considered as a cluster.

2) ACOUSTIC MODEL:

We used the acoustic models of the HUB4 BN EVAL-98 baseline system. Supervised adaptation of the acoustic models to lectures data is performed using the procedure described in [1]. For supervised adaptation, about 8h of speech data from the training partition of the TED corpus are used.

The adaptation procedure is the same as for the TC-STAR-P BN system: In contrast to the application of MLLR [2] in recognition, much more regression classes are used. A regression class tree is generated with an agglomerative clustering procedure [3]. Mean vectors are adapted using full transformation matrices, while diagonal transformation matrices are used to adapt the variances.

3) LANGUAGE MODEL:

The language model for the TED corpus is built by mixing training data from conference proceedings, lecture transcripts, and conversational speech transcripts:

- Lect 55Kw of lecture transcripts from the TED training data;
- Proc 15Mw of scientific papers from speech conferences and workshops (Eurospeech, ICASSP, ICSLP, etc.);
- Conv 300Kw of transcripts of conversational speech (Verbmobil, HUB5).
- The training data of the HUB4 LM

On some system configuration, Language model adaptation exploited the paper presented in each lecture, i.e. for each lecture a different adapted language model was used in recognition. Details on language model adaptation can be found in [1].

4) RECOGNITION LEXICON DESCRIPTION:

The same recognition lexicon as for the HUB4 BN EVAL-98 baseline system is used.

5) RECOGNITION PROCESS

The recognition process is the same as for the HUB4 BN EVAL-98 baseline system. The reported results are the output of the second pass of the recognizer using MLLR-adapted [2] models.

6) EXECUTION TIME:

The total execution time of the first and second decoding pass is about 24xRT on an Intel Xeon 2.4Ghz machine with 512KB cache and 4GB memory.

7) SUBMISSIONS SUMMARY:

For this task, we submit four results:

(B1) baseline: output of the first recognition step of the Hub4 system, without any acoustic or language model adaptation.

(B2) baseline+mllr: output of the second recognition step of the previous system, after three iterations of cluster-based unsupervised MLLR adaptation. Supervision provided by the output of system (B1).

(A2) adapted: output of the second recognition step, after three iterations of cluster-based unsupervised MLLR adaptation, of a system using acoustic models adapted on the TED training set. The Language Model is fixed, and its training data include also task-related data, as explained in sec. 3. Supervision provided by the output of the first step of the same system.

(L2) adapted-slm: same system as (A2). In this case, however, a specific LM was estimated and applied for each different lecture, exploiting the associated paper. Supervision provided by the output of the first step of the same system.

8) REFERENCES:

[1] M. Cettolo, F. Brugnara and M. Federico, “Advances in the Automatic Transcription of Lectures,“, ICASSP 2004, Montreal, 2004.

[2] C.J. Leggetter, P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models“, Computer Speech & Language, Vol. 9, pp. 171-185, 1995.

[3] N. Bertoldi, F. Brugnara, M. Cettolo, M. Federico and D. Giuliani, “Cross-task portability of a broadcast news speech recognition system“, Speech Communication, Vol. 38, pp. 335-347, 2002.

6.4 English BN Recognizer for Task 2 (TCStar_P) by ITC-IRST

ITC-irst speech recognition system

TC-STAR-P BROADCAST NEWS ENGLISH TEST CONDITION

1) ACOUSTIC FRONT-END

The acoustic front-end of the system that has been used for the TC-STAR-P BN corpus is the same as for the system employed for HUB4 BN EVAL-98 test condition, i.e. input speech is automatically segmented and clustered, the features are Mel-frequency Cepstral coefficients which have been normalized using Cluster-based Cepstral Mean and Variance Normalization (CMVN).

For adaptation and recognition we used the (automatically generated) segmentation and clustering of the TC-STAR-P corpus that has been provided by LIMSI.

2) ACOUSTIC MODEL:

We used the acoustic models of the HUB4 BN EVAL-98 baseline system. Results are reported for two systems:

(i) acoustic models of the HUB4 BN EVAL-98 baseline system are immediately applied to the TCSTAR-P BN corpus without adaptation on the training partition of this data set.

(ii) Supervised adaptation to the TCSTAR-P BN corpus is performed with the following steps:

a) words that are contained in the word transcriptions of the TCSTAR-P BN training subset but not in the recognition lexicon are phonetically transcribed.

As not all missing words could be transcribed and several segments are not marked in the LIMSI clustering file (i.e. they do not contain usable speech data) this yielded an amount of about 6.5 hours of usable speech data for adaptation.

b) use MLLR [1] for supervised adaptation of the acoustic models. In contrast to the application of MLLR in recognition, much more regression classes are used (about 2200). A regression class tree is generated with an agglomerative clustering procedure [2]. The regression class tree is built in two steps:
   - firstly, for each phoneme state, Gaussian components are hierarchically clustered;
   - secondly, the roots of trees obtained with the first step are clustered.

Base regression classes, corresponding to the leaves of the regression class tree, are formed by single Gaussian components. During adaptation a minimum

class occupancy count of 1000 is imposed. Mean vectors are adapted using full transformation matrices, while diagonal transformation matrices are used to adapt the variances.

3) LANGUAGE MODEL:

The same language model as for the HUB4 BN EVAL-98 baseline system is used.

4) RECOGNITION LEXICON DESCRIPTION:

The same recognition lexicon as for the HUB4 BN EVAL-98 baseline system is used. The lexicon entries that have been added recently for adaptation purposes (see section (2)) are not contained in the recognition lexicon.

5) RECOGNITION PROCESS

The recognition process is the same as for the HUB4 BN EVAL-98 baseline system. The reported results for systems (i) and (ii) (see section (2)) are the output of the second pass of the recognizer after three steps of cluster-based unsupervised MLLR adaptation [1].

6) EXECUTION TIME:

The execution time of the first and second decoding pass is 172738.34 seconds (92926.27 seconds + 79812.07 seconds) on an Intel Xeon 2.4Ghz machine with 512KB cache and 4GB memory. This corresponds to 33.9xRT.

7) SUBMISSIONS SUMMARY:

For this task, we submit three results:

(B1) baseline: output of the first recognition step of the Hub4 system, without any acoustic or language model adaptation.

(B2) baseline+mllr: output of the second recognition step of the previous system, after three iterations of unsupervised MLLR adaptation. Supervision provided by the output of system (B1).

(A2) adapted+mllr: output of the second recognition step, after three iterations of cluster-based unsupervised MLLR adaptation, of a system which uses acoustic models adapted on the TCSTAR_P EN training set. Supervision provided by the output of the first step of the same system.

8) REFERENCES:

[1] C.J. Leggetter, P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models“, Computer Speech & Language, Vol. 9, pp. 171-185, 1995.

[2] N. Bertoldi, F. Brugnara, M. Cettolo, M. Federico and D. Giuliani, “Cross-task portability of a broadcast news speech recognition system“, Speech Communication, Vol. 38, pp. 335-347, 2002.

6.5 English BN Recognizer for Task 3 (EPPS) by ITC-IRST

ITC-irst speech recognition system

TC-STAR PARLIAMENT TEST

1) ACOUSTIC FRONT-END

The acoustic front-end of the system that has been used for this task is the same as for the system employed for HUB4 BN EVAL-98 test condition, i.e. input speech is automatically segmented and clustered, the features are Mel-frequency Cepstral coefficients which have been normalized using Cluster-based Cepstral Mean and Variance Normalization (CMVN).

For recognition we used the manual segmentation provided by RWTH.

2) ACOUSTIC MODEL:

We used the acoustic models of the HUB4 BN EVAL-98 baseline system.

3) LANGUAGE MODEL:

The same language model as for the HUB4 BN EVAL-98 baseline system is used.

4) RECOGNITION LEXICON DESCRIPTION:

The same recognition lexicon as for the HUB4 BN EVAL-98 baseline system is used.

5) RECOGNITION PROCESS

The recognition process is the same as for the HUB4 BN EVAL-98 baseline system. In this case, however, the manual segmentation was used. Clusters were automatically generated by applying an agglomerative clustering procedure to all segments within the same class. As class labels, we used the combination of <condition>,<gender> provided in the manual transcriptions.

6) EXECUTION TIME:

The total execution time of the first and second decoding pass is about 34xRT on an Intel Xeon 2.4Ghz machine with 512KB cache and 4GB memory.

7) SUBMISSIONS SUMMARY:

For this task, we submit two results:

(B1) baseline: output of the first recognition step of the Hub4 system, without any acoustic or language model adaptation.

(B2) baseline+mllr: output of the second recognition step of the previous system, after three iterations of cluster-based unsupervised MLLR adaptation. Supervision provided by the output of system (B1).

8) REFERENCES:

[1] C.J. Leggetter, P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models“, Computer Speech & Language, Vol. 9, pp. 171-185, 1995.

[2] N. Bertoldi, F. Brugnara, M. Cettolo, M. Federico and D. Giuliani, “Cross-task portability of a broadcast news speech recognition system“, Speech Communication, Vol. 38, pp. 335-347, 2002.

6.6 English BN Recognizer by LIMSI

LIMSI English baseline system

All LIMSI baseline results for English (including the TC-STAR_P data and the EPPS data) were obtained using the version V1.3 of the LIMSI American-English broadcast news transcription system which is essentially a packaged version of the LIMSI RT02 system with updated language models. This is not the current best LIMSI system but it is an off the shelf system that has been extensively used over the past 2 years.

1) GENERAL SYSTEM DESCRIPTION

The LIMSI segmentation and clustering is based on an audio stream mixture model [4,6]. First, the non-speech segments are detected and rejected using GMMs representing speech, speech over music, noisy speech, pure-music and other background conditions. An iterative maximum likelihood segmentation/clustering procedure is then applied to the speech segments. The result of the procedure is a sequence of non-overlapping segments with their associated segment cluster labels. Each segment cluster is assumed to represent one speaker in a particular acoustic environment and is modeled by a GMM. The objective function is the GMM log-likelihood penalized by the number of segments and the number of clusters, appropriately weighted. Four sets of GMMs are then used to identify telephone segments and the speaker gender. Segments longer than 30s are chopped into smaller pieces by locating the most probable pause within 15s to 30s from the previous cut.

The speech recognizer [1,6] uses 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band (or 0-3.5kHz for telephone data) every 10ms. For each 30ms frame the Mel scale power spectrum is computed, and the cubic root taken followed by an inverse Fourier transform. Then LPC-based cepstrum coefficients are computed. The cepstral coefficients are normalized on a segment-cluster basis using cepstral mean removal and variance normalization. Thus each cepstral coefficient for each cluster has a zero mean and unity variance. The 39-component acoustic feature vector consists of 12 cepstrum coefficients and the log energy, along with the first and second order derivatives.

Each phone model is a tied-state left-to-right CD-HMM with Gaussian mixtures. The triphone-based context-dependent phone models are word-independent but position-dependent. The tied states are obtained by means of a decision tree. Word recognition is performed in three steps: 1) initial hypothesis generation, 2) word graph generation, 3) final hypothesis generation.

Step 1: Initial Hypothesis Generation - This step generates initial hypotheses which are then used for cluster-based acoustic model adaptation. This is done via one pass (less than 1xRT [5]) cross-word trigram decoding with gender-specific sets of position-dependent triphones (6298 tied states) and a trigram language model (17M trigrams and 8M bigrams). Band-limited acoustic models are used for the telephone speech segments.

Step 2: Word Graph Generation - Unsupervised acoustic model adaptation is performed for each segment cluster using the MLLR technique [2]. A word graph is generated for each segment in a one pass trigram decoding using position-dependent triphones with 11730 tied states (16 Gaussians) and the trigram used in step 1.

Step 3: Final Hypothesis Generation - The final hypothesis is generated after a second MLLR adaptation using the word graphs, a 4-gram model and a 32-Gaussian version of the acoustic models used in step 2.

2) ACOUSTIC TRAINING

The acoustic models were trained on about 150 hours of American-English broadcast new data, including the 1995, 1996, and 1997 official NIST Hub4 training data. The acoustic models are position-dependent triphones with tied states, obtained using a divisive decision tree based clustering algorithm. Two sets of gender-dependent acoustic models were built using MAP [3] adaptation of SI seed models for each of wideband and telephone band speech. The Hub4 training data was also used to build the Gaussian mixture models for sex identification, and music and telephone segment detection. About 2 hours of pure music portions of the acoustic training data were used to estimate the music GMM.

3) LANGUAGE MODEL TRAINING

The 2 language models (3-gram and 4-gram) were obtained by interpolation of backoff n-gram language models trained on the following data sets (through November 1998):

1- BN transcriptions from LDC (years 92-95) and from PSMedia (years 96 and 97, and Jan-Nov 1998)

2- All newspaper and newswire texts distributed by LDC (Jan'94 - Jun'98)

3- Transcriptions of the acoustic data, BN data (including the 1995 MarketPlace data), plus all the dev and test sets predating Dec'98.

The interpolation coefficients were chosen in order to minimize the perplexity on the 1999 evaluation data, and a set aside portion of development texts. The backoff LMs are derived from this interpolation by merging the 3 LM components. The word list contains 64906 words, selected to minimize the OOV rate on a set of development texts taken from June 1998.

4) RECOGNITION LEXICON DESCRIPTION

Pronunciations are based on a 48 phone set (3 of them are used for silence, filler words, and breath noises). A pronunciation graph is associated with each word so as to allow

for alternate pronunciations, including optional phones. The 65k vocabulary contains 64906 words including 72627 phone transcriptions. Frequent inflected forms have been verified to provide more systematic pronunciations. As done in the past, compound words for about 300 frequent word sequences subject to reduced pronunciations were included in the lexicon as well as the representation of frequent acronyms as words.

5) REFERENCES

[1] J.L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, "Transcription of Broadcast News", EuroSpeech, Sep. 1997.

[2] C.J. Legetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech & Language, Vol. 9, pp. 171-185, 1995.

[3] J.L. Gauvain and C.H. Lee, "Maximum A Posteriori for Multivariate Gaussian Mixture Observation of Markov Chains", IEEE Trans. on Speech and Audio Processing, pp. 291-298, 1994.

[4] J.L. Gauvain, L. Lamel, G. Adda, "Partitioning and Transcription of Broadcast News Data", Proc. ICSLP'98, pp. 1335-1338, Sydney, Australia, December 1998.

[5] J.L. Gauvain, L. Lamel, "Fast Decoding for Indexation of Broadcast Data", ICSLP'2000, vol. 3, pp. 794-798, Beijing, Oct. 2000.

[6] J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI broadcast news transcription system", Speech Communication, vol. 37(1-2), pp. 89-108, May 2002.

6.7 English BN Recognizer by RWTH

English TC-STAR
TASK 1 (HUB4 BASELINE) - TASK 2 (TCSTAR-P) - TASK 3 (EPPS)
RWTH

0) INTRODUCTION

The baseline system used for the three evaluation tasks is essentially the same as the one described in [2]. For task 1, the system was run on the DARPA BN Eval98 set (LDC2000S86). For task 2, the baseline number is provided for the TCSTAR-P BN evaluation set, using the LIMSI partitioning. For task 3, the baseline number for the one hour TCSTAR parliamentary speech is provided, using the manual segmentation provided with the data.

Recognition was performed in a single pass without MLLR and VTN, and gender-dependent acoustic models were used.

1) ACOUSTIC ANALYSIS

We used standard MFCC features.
The magnitude spectrum was estimated by applying the DFT to the preemphasised and windowed audio signal each 10ms. Next the magnitude spectrum was filtered with a filter bank consisting of 20 triangular filters positioned at equidistant points on the Mel frequency axis. The logarithms of the filter outputs were cepstrally decorrelated (discrete cosine transform), resulting in 16 dimensional vectors. The MFCCs were normalised using cepstral mean removal, and energy and variance normalisation. Nine temporally consecutive vectors were fed into an LDA to obtain 45 dimensional feature vectors which were used for the baseline results.

2) ACOUSTIC MODEL

The words of the vocabulary were modelled by position-dependent [2] triphones with across-word contexts [1]. The triphones were modelled by Hidden Markov Models (HMMs). The non-silence HMMs are standard three-states left-to-right HMMs, whereas the silence HMM consists of a single HMM state. The emission probabilities assigned to the HMM states in turn were modelled by gender-dependent Gaussian mixture models, sharing a single, globally pooled diagonal covariance matrix. The transition probabilities were set empirically. A gender-dependent binary decision tree (CART) with 137 questions was used to tie the HMM states. During training and recognition we applied the Viterbi approximation.

For tasks 1, 2, and 3, the tied states were trained on data out of the HUB4 1996 (LDC97S44) and HUB4 1997 (LDC98S71) training corpora which were manually

checked for transcription errors. The training material summed up to 96 hours. The acoustic model consists of 4001 tied states. The acoustic model consists of 200.500 (female) and 350.000 (male) densities.

3) LANGUAGE MODEL

For tasks 1, 2, and 3 a conventional HUB4 trigram language model was used, cf. [2]. The oov-rate on task 1 was 1.3%, on task 2 5.5%, and on task 3 1.6%. The numbers were achieved on the HUB4EN 98 development set, on the TCSTAR-P development set, and on the EPPS test set, respectively.

4) RECOGNITION LEXICON

The lexicon for tasks 1, 2, and 3 is identical to the one used in [2]. It contains a single phoneme modelling filled pauses. During recognition all noise events were mapped to silence. The lexicon considers pronunciation variants and phrases. Finally, the lexicon consists of 43 non-silence phonemes, 66,272 words, and 77,638 pronunciations.

5) RECOGNITION

Our baseline system was a gender-dependent, single pass across-word recogniser. A beam search strategy with a pre-pruning step based on language model look-ahead using a bigram model [2] was applied. Neither VTN nor MLLR were used to produce the baseline results.

6) EXECUTION TIME

On an AMD Athlon MP with 1800Mhz and 3GB RAM a real-time factor of about 10 was measured for all tasks.

7) REFERENCES

[1] A. Sixtus, S. Molau, S. Kanthak, R. Schlüter, H. Ney. “Recent Improvements of the RWTH Large Vocabulary Speech Recognition System on Spontaneous Speech“. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1671-1674, Istanbul, Turkey, June 2000.

[2] A. Sixtus. “Across-Word Phoneme Models for Large Vocabulary Continuous Speech Recognition“. Dissertation, Aachen, Germany, January 2003.

6.8 English BN Recognizer by Sony

Primary Test System Description

The SONY BN baseline speech recognizer uses 32 LDA coefficients computed from a 38-dimensional source feature space. For the 38 source parameters, 12 Mel frequency cepstrum coefficients, along with the first and second order derivatives and first and second order derivatives of the log energy are used.

The Mel frequency power spectrum is estimated on the 0Hz-8kHz band every 10ms. For each 16ms frame, the signal is first preemphasized with a first order IIR filter (1/(1+0.97*z-1)), a Hamming Window is applied and the Mel-scaled FFT power spectrum is computed. The frequency band used by the triangular filter bank is finally 80Hz-7500Hz.

The cepstral coefficients are normalized on a segment-cluster basis using cepstral mean removal and variance normalization.

Each phone model is a tied-state left-to-right CD-HMM with Gaussian mixtures. The triphone-based context-dependent phone models are word-independent but position-dependent. The tied states are obtained by means of a divisive decision tree based clustering algorithm.

The baseline system does not perform acoustic model adaptation (like e.g. MLLR) or feature space adaptation (except for the LDA).

Word recognition is performed in 4 steps:
1) Tree Pass
    Time synchronous search without word copies on a tree-organized dictionary
2) Flat Pass
    Time synchronous search on a rolled-out dictionary: poor-man's trigrams.
3) Lattice generation
4) Lattice rescoring using the full trigram

Acoustic Training

The acoustic models were trained on 96206 utterances taken from an internal Sony spontaneous speech database (77387 utterances) and from the LDC 1996 English Broadcast News Speech data LDC97S44 (18819 utterances, total of 104 hours of broadcasts from ABC, CNN and CSPAN television networks and NPR and PRI radio networks with corresponding transcripts). The training is done over 5 iterations of Viterbi alignment.

The acoustic models are position-dependent triphones with tied states, obtained using a divisive decision tree based clustering algorithm.
Grammar Training

The n-gram language models were obtained by interpolation of back off n-gram language models trained on the following sources:

Corpora Name	Approx. # Words (Millions)
BN LDC98T31 (Broadcast News, Jan 1992 - Apr 1996)	130.29
WSJ (1987-1989)	36.81
WEB TEXT	13.74

The Interpolation weights were computed through EM optimization of PPL in the development test. The back-off weights are rescaled after interpolation. Pruned version: Pruning based on relative entropy. A 64K vocabulary is used. The vocabulary selection is based on thresholding for words on the BN acoustic text. Additional vocabulary is taken from most frequent words from BN LM text.

Number of N-grams: 1-gram=64001, 2-gram=7670581, 3-gram=8646829.

Recognition Lexicon Description

The pronunciations are based on a 50 phones set (6 of them are used for silence, filler words, and different noises). A pronunciation graph is associated with each word so as to allow for alternate pronunciations. The 24k vocabulary contains 25508 words with 51578 phones transcriptions.

Results

The following results were obtained with this Baseline with the NIST evaluation tools on the Hub4e_98 task for the two files h4e_98_1,sph and h4e_98_2.sph:

SPKR	Overall	Broadcast Baseline Speech (F0)	Spontaneous Broadcast Speech (F1)	Speech Over Telephone Channels (F2)	Speech in the Presence of Background Music (F3)	Speech Under Degraded Acoustic Conditions (F4)	Speech from Non-Native Speakers (F5)	All other speech (Fx)
Avg 1	47.2	37.0	42.7	61.2	54.5	44.9	70.9	57.8
Avg 2	47.7	42.5	43.2	54.3	44.6	54.5	64.3	68.2

6.9 English BN Recognizer for Task 1 (TED) by UKA

UNIVERSITÄT KARLSRUHE (TH), INTERACTIVE SYSTEMS LABORATORIES
THE ISL BASELINE LECTURE TRANSCRIPTION SYSTEM FOR THE TED CORPUS
TASK 1; Translanguage English Database (TED)
MANUAL SEGMENTATION

1) PRIMARY TEST SYSTEM DESCRIPTION:

Our recognition experiments were conducted on the *Translanguage English Database* (TED) corpus [1] which is a corpus of recordings made of oral presentations at Eurospeech 1993 in Berlin. The chosen material is challenging on several aspects: lecture speech varies in speaking style from freely presented to read, comprising spontaneous events as well as hyper articulation. The TED corpus contains mainly non-native speakers of English, some not even fluent. The recorded audio files vary in quality and are partially noisy. Our test set contained the same eight speakers as published by IRST [2] (6 male speakers, Sp.4 and Sp.5 female) with a wide variety of mother tongues (Sp.1: English, Sp.2: Italian, Sp.3: French, Sp.4: French, Sp.5: Danish, Sp.6: German, Sp.7: Dutch, Sp.8: Japanese).

The ISL Baseline Lecture Transcription System for the TED Corpus is similar to the ISL RT04S Meeting Transcription System described in [3], using similar acoustic model, but a different dictionary and language model. The segmentation provided by the manual transcription of the test corpus was used as given, without any modifications. Generic speakers were not clustered across speeches.

The decoding process takes place in two stages of subsequent systems that are adapted on the hypotheses of the previous system. A description of the system is given in section 2 to section 4.

Step 1:
In this step a first set of hypotheses is generated using a simple system not further described.

Step 2:
Warping factors for *vocal track length normalization* (VTLN) are estimated. Then the acoustic is adapted using *maximum likelihood linear regression* (MLLR) and *feature space adaptation* (FSA). The final set of hypotheses is created.

All decoding stages consist of a single run with our IBIS single pass decoder [4] generating a word lattice, a confusion network, and a lattice rescoring using a different set of language model parameters (language model weight and word penalty).

2) ACOUSTIC MODEL TRAINING:

The speech recognition experiments described below were conducted with the *Janus Recognition Toolkit* (JRTk), which was developed and is maintained jointly by the Interactive Systems Laboratories at the Universität Karlsruhe (TH), Germany and at the Carnegie Mellon University in Pittsburgh, USA.

As relatively little supervised data is available for acoustic modelling of the TED corpus the acoustic model has been trained on *Broadcast News* [5] and merged with the close talking channel of meeting corpora [3] [6] summing up to a total of 300 hours of training material.

The speech data was sampled at 16kHz. Speech frames were calculated using a 10ms Hamming window. For each frame, 13 *Mel-Minimum Variance Distortionless Response* (Mel-MVDR) cepstral coefficients were obtained through a discrete cosine transform from the Mel-MVDR spectral envelope [7]. Thereafter, linear discriminant analysis was used to reduce the utterance based cepstral mean normalized features plus 7 adjacent to a final feature number of 42. Our baseline model consisted of 300k Gaussians with diagonal covariances organized in 24k distributions over 6k codebooks.

2) ACOUSTIC MODEL ADAPTATION

The adaptation of the acoustic model was obtained by three consecutive steps:

Step 1:
A supervised Viterbi training of the TED adaptation speakers followed by a *maximum a posteriori* (MAP) combination of this model with the acoustic model of the original system: To find the best mixing weight, a grid search over different mixing weights was performed.
The weight, which reached the best likelihood on the hypotheses of the first pass of the unadapted speech recognition system, was chosen as the
final mixing weight.

Step 2:
A supervised *maximum likelihood linear regression* (MLLR) [8] in combination with *feature space adaptation* (FSA) and *vocal track length normalization* (VTLN) on the TED adaptation speakers: This step adapts to the speaking style of the lectures and the channel.

Step 3:
A second, now unsupervised MLLR, FSA and VTLN adaptation based on the hypothesis of the first recognition run: this procedure aims at adapting to the particular speaking style of a speaker and to changes within the channel.

4) GRAMMAR TRAINING:
To generate *language models* (LM) for interpolation we used corpora consisting of broadcast news (160M words), *proceedings* (17M words) of conferences such as ICSLP, Eurospeech, ICASSP or ASRU and *talks* (60k words) by the TED adaptation

speakers. Our final LM was generated by interpolating a 3-gram LM based on broadcast news and proceedings, a class based 5-gram LM based on broadcast news and proceedings and a 3-gram LM based on the talks. The usage of student presentations about speech related topics recorded at the Universität Karlsruhe (TH) as well as conversational speech such as the Verbmobil corpus was not helpful in decreasing the perplexity and the WER. The overall out of vocabulary rate is 0.3% by a vocabulary size of 25,000 words including multi-words and pronunciation variants.

4) PERFORMANCE:

The system achieved a word error rate of 31.0%.

5) REFERENCES:

[1] Linguistic Data Consortium (LDC), “Translanguage English Database”, www.ldc.upenn.edu/Catalog/LDC2002S04.html.

[2] E. Leeuwis, M. Federico, and M. Cettolo, “Language modeling and transcription of the TED corpus lectures“, *ICSSP*, 2003.

[3] F. Metze, Q. Jin, C. Fügen, K. Laskowski, Y. Pan, and T. Schultz “Issues in Meeting Transcription – The ISL Meeting Transcription System”, in Proc. ICSLP 2004. Jeju Island, Korea

[4] 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, 12-2001

[5] Linguistic Data Consortium (LDC), “English Broadcast News Speech (Hub-4)” www.ldc.upenn.edu/Catalog/LDC97S44.html.

[6] S. Burger, V. Maclaren, and H. Yu, “The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style“, *ICSLP*, 2002.

[7] M.C.Wölfel, J.W. McDonough, and A.Waibel, “Warping and Scaling of the Minimum Variance Distortionless Response“, *ASRU*, 2003.

[8] C. J. Leggetter and P. C. Woodland, “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models“, *Computer Speech and Language*, pp. 171–185, 1995.

6.10 English BN Recognizer for Task 2 (TCStar_P) by UKA

UNIVERSITÄT KARLSRUHE (TH), INTERACTIVE SYSTEMS LABORATORIES
ISL BCAST1EN BASELINE TRANSCRIPTION SYSTEM
TASK 2; TC-STAR-P, EUROPEAN ENGLISH BROADCAST NEWS (BN)
AUTOMATIC SEGMENTATION PROVIDED BY LIMSI

1) PRIMARY TEST SYSTEM DESCRIPTION:

The ISL BCAST1EN Baseline Transcription System is fundamentally the same as the ISL RT04S Meeting Transcription System described in [1], using the same acoustic model, dictionary, and language model, trained for transcribing meetings held in English.
The automatic segmentation was provided by LIMSI and used without any modifications.

For cross-adaptation purposes we also made use of two sets of acoustic models coming from the 2003 ISL Rich Transcription System for Conversational Telephony Speech (Switchboard) [2].

The decoding process takes place in five stages of subsequent systems that are adapted on the hypotheses of the previous system. A description of the different systems is given in section 2 as well as in [1] and [2].

Step 1: Initial Hypothesis Generation. In this step a first set of hypotheses is generated using the system PLAIN, no adaptation being performed.

Step 2: Tree6. Warping factors for VTLN are estimated using the Tree6 Switchboard acoustic. Then the Tree6 acoustic is adapted using MLLR and feature-space constrained MLLR (CMLLR). Using the adapted system a second set of hypotheses is generated.

Step 3: Tree150. VTLN warping factors are re-estimated. The Tree150 Switchboard acoustic is adapted using MLLR and CMLLR. A third set of hypotheses is generated.

Step 4: MAS. The MT MAS acoustic is being adapted as described in the previous stages, a fourth set of hypotheses is generated.

Step 5: SAT. The MT SAT acoustic is adapted as above, the final set of hypotheses is created.

All decoding stages consist of a single run with our IBIS single pass decoder [3] generating a word lattice, and a lattice rescoring using a different set of language model parameters (language model weight and word penalty).

Unlike in the RT04S Meeting Transcription System no confusion network combination was performed.

2) ACOUSTIC TRAINING:

The acoustic models that were developed specifically for the meeting task transcription system (PLAIN, MAS, SAT) were trained on 180 hours of Broadcast News training data from 1996 and 1997, as well as meeting data from three different sites:

CMU: 11h
ICSI: 72h
NIST: 13h

A detailed description of the training corpora can be found in [4,5,6,7,8]

Using the above data, an acoustic model using ~300k Gaussians with diagonal covariances organized in 24k distributions over 6k codebooks in a 42-dimensional feature space based on MFCCs after LDA with utterance-based cepstral mean subtraction was trained. All systems except the SAT system made use of a global STC transformation matrix after LDA.

The acoustic models take from the ISL Switchboard (SWB) system (Tree6, Tree150) were trained on a merger of 265h of SWB and Callhome, 32h of cellphone and 65h of "CTRAN" SWB-2 data. The acoustic preprocessing is based on 13 MFCC per frame, speaker wide cepstral mean subtraction, concatenation of 11 frames, using LDA to reduce the dimension of the feature vector to 42. The acoustic model is organized in 50k distributions making use of 10k codebooks. Two different kinds of clustering trees were trained. The Tree150 system uses a clustering scheme with one sub-tree for each context-independent HMM sub-state. Tree6 utilizes a clustering tree consisting of only six sub-trees, allowing cross-phone sharing of parameters.

Overview of the systems used:

PLAIN: Merge-and-split training followed by 2 iterations of viterbi training, trained on close talking data, not VTLN [1]

Tree6: Tree6 Switchboard acoustic, using cross-phone parameter sharing, merge-and-split training [2]

Tree150: Tree 150 Switchboard acoustic, using our traditional clustering scheme, merge-and-split training [2]

MAS: Merge-and-split training followed by 2 iterations of viterbi training, VTLN, 6000 codebooks, 24000 distributions [1]

SAT: Speaker adaptive training on close-talking microphone data, no STC, 6000 codebooks, 24000 distributions; this system was not used in the RT04S NIST evaluation.

3) GRAMMAR TRAINING:
Language Models were trained in analogy to the Switchboard system. We trained a simple 3-gram LM and a 5-gram LM with ~800 automatically introduced 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 ~47k words.

4) PERFORMANCE:
The system achieved a word error rate of 46.7% on the official TC-STAR task 2 test corpus for English.

5) REFERENCES:

[1] F. Metze, Q. Jin, C. Fügen, K. Laskowski, Y. Pan, and T. Schultz “Issues in Meeting Transcription – The ISL Meeting Transcription System”, in Proc. ICSLP 2004. Jeju Island, Korea

[2] H. Soltau, H. Yu, F. Metze. C. Fügen, Q. Jin, and S.-C. Jou, “The 2003 ISL Rich Transcription System for Conversational Telephony Speech“, in Proc. ICASSP 2004. Montreal; Canada: IEEE 2004

[3] 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, 12-2001

[4] S.Burger, V. MacLaren, and H. Yu, “The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style,” in Proc. ICSLP 2002. Denver, CO: ISCA, 9 2002

[5] S.Burger and Z. Sloan, “The ISL Meeting Corpus: Categorical Features of Communicative Group Interactions,” in Proc. ICASSP-2004 Meeting Recognition Workshop. Montreal; Canada: NIST, 5 2004.

[6] A. Janin, J. Ang, S.Bhagat, R. Dhillon, J. Edwards, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede, “The ICSI Meeting Project: Resources and Research,” in Proc. ICASSP-2004 Meeting Recognition Workshop. Montreal; Canada: NIST, 5 2004.

[7] S. Strassel and M. Glenn, “Shared Linguistic Resources for Human Language Technology in the Meeting Domain,” in Proc. ICASSP-2004 Meeting Recognition Workshop. Montreal; Canada: NIST, 5 2004.

[8] V. Stanford and J. Garofolo, “Beyond Close-talk – Issues in Distant Speech Acquisition, Conditioning Classification, and Recognition,” in Proc. ICASSP-2004 Meeting Recognition Workshop. Montreal; Canada: NIST, 5 2004.

6.11 English BN Recognizer for Task 3 (EPPS) by UKA

UNIVERSITÄT KARLSRUHE (TH), INTERACTIVE SYSTEMS LABORATORIES
ISL EPPS BASELINE TRANSCRIPTION SYSTEM
TASK 3; EUROPEAN PARLIAMENTARY PLENARY SPEECHES (EPPS)
MANUAL SEGMENTATION

1) PRIMARY TEST SYSTEM DESCRIPTION:

The ISL EPPS Baseline Transcription System is fundamentally the same as the ISL RT04S Meeting Transcription System described in [1], using the same acoustic model, dictionary, and language model, trained for transcribing meetings held in English.

The segmentation provided by the manual transcription of the test corpus was used as given, without any modifications. Generic speakers, such as interpreters, that were not identified by name were not clustered across speeches.

For cross-adaptation purposes we also made use of two sets of acoustic models coming from the 2003 ISL Rich Transcription System for Conversational Telephony Speech (Switchboard) [2].

The decoding process takes place in five stages of subsequent systems that are adapted on the hypotheses of the previous system. A description of the different systems is given in section 2 as well as in [1] and [2].

Step 1: Initial Hypothesis Generation. In this step a first set of hypotheses is generated using the system PLAIN, no adaptation being performed.

Step 2: Tree6. Warping factors for VTLN are estimated using the Tree6 Switchboard acoustic. Then the Tree6 acoustic is adapted using MLLR and feature-space constrained MLLR (CMLLR). Using the adapted system a second set of hypotheses is generated.

Step 3: Tree150. VTLN warping factors are re-estimated. The Tree150 Switchboard acoustic is adapted using MLLR and CMLLR. A third set of hypotheses is generated.

Step 4: MAS. The MT MAS acoustic is being adapted as described in the previous stages, a fourth set of hypotheses is generated.

Step 5: SAT. The MT SAT acoustic is adapted as above, the final set of hypotheses is created.

All decoding stages consist of a single run with our IBIS single pass decoder [3] generating a word lattice, and a lattice rescoring using a different set of language model parameters (language model weight and word penalty).

Unlike in the RT04S Meeting Transcription System no confusion network combination was performed.

2) ACOUSTIC TRAINING:

The acoustic models that were developed specifically for the meeting task transcription system (PLAIN, MAS, SAT) were trained on 180 hours of Broadcast News training data from 1996 and 1997, as well as meeting data from three different sites:

CMU: 11h
ICSI: 72h
NIST: 13h

A detailed description of the training corpora can be found in [4,5,6,7,8]

Using the above data, an acoustic model using ~300k Gaussians with diagonal covariances organized in 24k distributions over 6k codebooks in a 42-dimensional feature space based on MFCCs after LDA with utterance-based cepstral mean subtraction was trained. All systems except the SAT system made use of a global STC transformation matrix after LDA.

The acoustic models take from the ISL Switchboard (SWB) system (Tree6, Tree150) were trained on a merger of 265h of SWB and Callhome, 32h of cellphone and 65h of "CTRAN" SWB-2 data. The acoustic preprocessing is based on 13 MFCC per frame, speaker wide cepstral mean subtraction, concatenation of 11 frames, using LDA to reduce the dimension of the feature vector to 42. The acoustic model is organized in 50k distributions making use of 10k codebooks. Two different kinds of clustering trees were trained. The Tree150 system uses a clustering scheme with one sub-tree for each context-independent HMM sub-state. Tree6 utilizes a clustering tree consisting of only six sub-trees, allowing cross-phone sharing of parameters.

Overview of the systems used:

PLAIN: Merge-and-split training followed by 2 iterations of viterbi training, trained on close talking data, not VTLN [1]

Tree6: Tree6 Switchboard acoustic, using cross-phone parameter sharing, merge-and-split training [2]

Tree150: Tree 150 Switchboard acoustic, using our traditional clustering scheme, merge-and-split training [2]

MAS: Merge-and-split training followed by 2 iterations of viterbi training, VTLN, 6000 codebooks, 24000 distributions [1]

SAT: Speaker adaptive training on close-talking microphone data, no STC, 6000 codebooks, 24000 distributions; this system was not used in the RT04S NIST evaluation.

3) GRAMMAR TRAINING:

Language Models were trained in analogy to the Switchboard system. We trained a simple 3-gram LM and a 5-gram LM with ~800 automatically introduced 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 ~47k words.

4) PERFORMANCE:

The system achieved a word error rate of 32.0% on the official TC-STAR task3 test corpus for English, 03may2004 EPPS.

5) EXECUTION TIME:

The system's total execution time was 251015 seconds, measured on a PC running SuSE Linux equipped with a Pentium 4 3.00 GHz, and 2 GB of RAM. Since the length of the test set that was decoded is 3600 seconds this results in a real-time factor (RTF) of 69.7

The run time of the single steps was as follows:

Step1:
Decoding: 42213s

Step2:
Adaptation: 3897s
Decoding: 30814s

Step3:
Adaptation: 4697s
Decoding: 59683s

Step4:
Adaptation: 5320s
Decoding: 56218s

Step5:
Adaptation: 3950s
Decoding: 44223s

Total: 251015s
Decoded Material: 3600s
RTF: 69.7

6) REFERENCES:

[1] F. Metze, Q. Jin, C. Fügen, K. Laskowski, Y. Pan, and T. Schultz “Issues in Meeting Transcription – The ISL Meeting Transcription System”, in Proc. ICSLP 2004. Jeju Island, Korea

[2] H. Soltau, H. Yu, F. Metze. C. Fügen, Q. Jin, and S.-C. Jou, “The 2003 ISL Rich Transcription System for Conversational Telephony Speech,“ in Proc. ICASSP 2004. Montreal; Canada: IEEE 2004

[3] 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, 12-2001

[4] S.Burger, V. MacLaren, and H. Yu, “The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style,” in Proc. ICSLP 2002. Denver, CO: ISCA, 9 2002

[5] S.Burger and Z. Sloan, “The ISL Meeting Corpus: Categorical Features of Communicative Group Interactions,” in Proc. ICASSP-2004 Meeting Recognition Workshop. Montreal; Canada: NIST, 5 2004.

[6] A. Janin, J. Ang, S.Bhagat, R. Dhillon, J. Edwards, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede, “The ICSI Meeting Project: Resources and Research,” in Proc. ICASSP-2004 Meeting Recognition Workshop. Montreal; Canada: NIST, 5 2004.

[7] S. Strassel and M. Glenn, “Shared Linguistic Resources for Human Language Technology in the Meeting Domain,” in Proc. ICASSP-2004 Meeting Recognition Workshop. Montreal; Canada: NIST, 5 2004.

[8] V. Stanford and J. Garofolo, “Beyond Close-talk – Issues in Distant Speech Acquisition, Conditioning Classification, and Recognition,” in Proc. ICASSP-2004 Meeting Recognition Workshop. Montreal; Canada: NIST, 5 2004.

7 System Descriptions Spanish

7.1 Spanish BN Recognizer by IBM

IBM Spanish TC-STAR Baseline System Description

1) PRIMARY TEST SYSTEM DESCRIPTION:

One single baseline system was run on the first 2 evaluation tasks. For task 1, baseline numbers are provided for the DARPA Broadcast News Hub-4NE Eval97 task, with automatic data partitioning. The Partitioning for this data was done with the acoustic segmentation software provided by CMU [1]. For task 2, baseline results are submitted for the TC_STAR_P BN evaluation set, using the short segments (pem4) created by automatic partitioning [2] provided by LIMSI.

The BN system uses 42 dimensional feature vectors obtained from an LDA projection. The source space for the LDA projection is 117 dimensional and obtained by stacking 9 temporally consecutive 13 dimensional acoustic observation vectors. The vectors contain 12 cepstral parameters obtained from an inverse DCT of the log outputs of a 24 band, triangular filter bank. The filters are positioned at equidistant points on the Mel-frequency scale between 0 and 8 kHz. In addition to the 12 cepstral parameters, the vectors contain a raw frame energy parameter. The ceptral parameters are mean normalized on a per utterance basis. The energy parameters are translated based on the max energy, also on a per utterance basis.

The system uses a single speaker independent acoustic model. The model is a continuous density left-to-right HMM using 16 component Gaussian mixture emission distributions and uniform transition probabilities. Each HMM has 3 states except for the silence HMM which is a single state model. The system uses 53 phones, 49 speech phones with stress markers, 1 silence phone, 1 speaker noise phone, 1 mumble phone and 1 filled pause phone. The speech HMMs use about 3000 context dependent tied state distributions obtained by decision tree clustering of triphone statistics using context questions based on 100 phonetic classes. In addition, each model uses a global Semi-Tied Covariance (STC)[3,4] linear transformation.

Decoding was performed in a single pass using a static decoding graph with 20M arcs and 15M states that was built from a trigram language model (47k 1-grams, 2.4M 2-grams and 2.5M 3-grams), a 47k word lexicon and the HMM components. The acoustic models use cross-word contexts.

2) ACOUSTIC TRAINING:

This IBM BN system for Spanish was trained from scratch for the TCStar Baseline Experiments with the suggested LDC Broadcast News data without using any bootstrapping system. The acoustic models were trained on a subset of the Hub-4NE 1997 (LDC98S74) training set, which consist of about 30h of recordings. A large development test set out of that data was hold back and all recordings with transcripts

not covered by the existing pronunciation dictionaries were dropped. Using the timing information in the corresponding transcripts (LDC98T29), this data was processed to also discard non-speech and overlapping speech segments and resulted in about 20 hours of usable recordings. Furthermore, the transcripts were text normalized to about 200K words (with about 16k unique words). Mispronounced words were modelled using the mumble phone.

3) GRAMMAR TRAINING:

The 3-gram language model was built using the SRI LM toolkit [5] with standard options and some pruning [6] in order to reduce bi and tri-grams. Training text was taken from the BN transcripts of the hub4 acoustic training data excluding a couple of shows that were hold back for testing. Additional text was taken from news articles, as given below:

Lines	Words	Source
30K	300K	Subset of Hub4 Training transcripts
1.5M	43M	El Pais
0.9M	25M	Heraldo de Aragon
6.3M	140M	LDC95T9 Spanish News Text
8.8M	210M	total

4) RECOGNITION LEXICON DESCRIPTION:

The 47k lexicon was obtained by taking words that occurred more than 100 times in one of the LM training corpora and adding all words of the acoustic training transcripts. Pronunciations are based on an IBM specific Spanish phone set with 49 speech phones including stressed phones augmented by 1 silence word and phone, as well as one for speaker noise and filled pauses. 34k words were covered by existing pronunciation dictionaries and the rest was created with an automatic phonetizer without any manual checking. It is clear that this text data with BN shows from before 1997 and news articles from before 1995 does not match very well the TCStar_P data dated in 2003.

5) EXECUTION TIME

As agreed among project partners no particular attention was spent on real time. Recognition was performed on a Linux cluster with different machines all based on Pentium 4 type processors with 2GB memory.

6) REFERENCES

[1] Siegler, M., Jain, U., Raj, B., Stern, R., “Automatic Segmentation, Classification and Clustering of Broadcast News Audio“, Proc. of the 1997 ARPA Speech Recognition Workshop, pp. 97-99, Feb. 1997.

[2] J. L. Gauvain, L. Lamel and G. Adda, “The LIMSI Broadcast News Transcription System“, Speech Communication, 37(1-2), pp. 89-108, 2002.

[3] M. J. F. Gales, “Maximum Likelihood Linear Transformations for HMM-based Speech Recognition“, in Computer Speech and Language, No. 12, pp. 75-98, 1998.

[4] G. Saon, G. Zweig and M. Padmanabhan, “Linear Feature Space Transformations for Speaker Adaptation“, in Proceedings International Conference on Acoustics Speech and Signal Processing, Salt Lake City, UT, 2001.

[5] A. Stolcke, “SRILM - An Extensible Language Modeling Toolkit“, in Proceedings International Conference on Spoken Language Processing, Denver, CO, Sept. 2002.

[6] A. Stolcke, “Entropy-based pruning of backoff language models“, in Proceedings DARPA Broadcast News Transcription and Understanding Workshop, pp. 270-274, Lansdowne, VA, Feb. 1998, Morgan Kauffmann.

7.2 Spanish BN Recognizer by LIMSI

LIMSI Spanish baseline system

1) GENERAL SYSTEM DESCRIPTION

The LIMSI 1x Spanish Broadcast News system has two main parts:
automatic partitioning and speech recognition.

The partitioning procedure is as follows [1,2]: First, the non-speech segments are detected (and rejected) using GMMs. Four GMMs each with 64 Gaussians serve to detect speech, pure-music and other (background). All test segments labeled as music or silence are removed prior to further processing. An iterative maximum likelihood segmentation/clustering procedure is then applied to the speech segments using GMMs and an agglomerative clustering algorithm. Given the sequence of cepstral vectors the algorithm tries to maximize an objective function defined as a penalized log-likelihood. Alternate Viterbi reestimation and agglomerative clustering gives a sequence of estimates with non-decreasing values of the objective function. The algorithm stops when no merge is possible. A constraint on the cluster size is used to ensure that each cluster corresponds to at least 10s of speech. This procedure is controlled by 3 parameters: the minimum cluster size (10s), the maximum log-likelihood loss for a merge, and the segment boundary penalty. When no more merges are possible, the segment boundaries are refined (within a 1s interval) using the last set of GMMs and an additional relative energy-based boundary penalty. This is done to locate the segment boundaries at silence portions, so as to avoid cutting words. Speaker-independent GMMs corresponding to wideband speech and telephone speech (each with 64 Gaussians) are then used to label telephone segments. This is followed by segment-based gender identification, using 2 sets of GMMs with 64 Gaussians (one for each bandwidth). The result of the partitioning process is a set of speech segments with cluster, gender and telephone/wideband labels.

The LIMSI BN speech recognizers [2] use 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band (or 0-3.5kHz for telephone data) every 10ms. For each 30ms frame the Mel scale power spectrum is computed, and the cubic root taken followed by an inverse Fourier transform. Then LPC-based cepstrum coefficients are computed. The cepstral coefficients are normalized on a segment-cluster basis using cepstral mean removal and variance normalization. Thus each cepstral coefficient for each cluster has a zero mean and unity variance. The 39-component acoustic feature vector consists of 12 cepstrum coefficients and the log energy, along with the first and second order derivatives.

Each phone model is a tied-state left-to-right CD-HMM with Gaussian mixtures. The triphone-based context-dependent phone models are word-independent but position-dependent. The tied states are obtained by means of a decision tree. A one-pass cross-word trigram decoding is carried out in about 0.5xRT using gender-specific sets of position-dependent triphones (1574 tied states) and a trigram language model (24M trigrams and 15M bigrams). Band-limited acoustic models are used for the telephone

speech segments. The 3-gram word lattice which is then expanded with a 4-gram LM. The posterior probabilities of the lattice edges are estimated using the forward-backward algorithm. The the 4-gram lattices are converted to a confusion network with posterior probabilities by iteratively merging lattice vertices and splitting lattices edges until a linear graph is obtained. This procedure gives comparable results to the edge clustering algorithm proposed in [3]. The words with the highest posterior in each confusion set are hypothesized.

2) ACOUSTIC TRAINING

The acoustic models were trained on most of the Hub-4NE 1997 (LDC98S74) training set. The acoustic models are position-dependent triphones with tied states, obtained using a divisive decision tree based clustering algorithm. Two sets of gender-dependent acoustic models were built using both MAP adaptation [4] of SI seed models for each of wideband and telephone band speech.

The English Hub4 training data was used to build the Gaussian mixture models for gender identification, and music and telephone segment detection. About 2 hours of pure music portions of the acoustic training data were used to estimate the music GMM.

3) LANGUAGE MODEL TRAINING

The n-gram language models were obtained by interpolation [5] backoff n-gram language models trained on the following data sets:

1- Hub-4NE 1997 transcriptions (1.9M words)

2- All newspaper and newswire texts distributed by LDC: 389M words

3- Articles from the online newspaper Caretas: 9.6M words

The Caretas source was used to have a more recent source for LM training data than the data distributed by the LDC.

The 65k word list was selected from the same text sources so as to minimize the OOV rate on the dev data. The word list contains 64999 words plus [silence], <s> and </s> and has an OOV rate of 1.4% on the eval97 test.

4) RECOGNITION LEXICON DESCRIPTION

Pronunciations are based on a 27 phone set (3 of them are used for silence, filler words, and breath noises). A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones. The 65k vocabulary contains 64999 words including 79156 phone transcriptions. The pronunciations were all generated automatically.

5) EXECUTION TIME

The execution time was not looked at closely, but it is around 0.8xRT for the NIST Eval97 test set and 2-3xRT for the TCStar_P test set, including segmentation.

6) REFERENCES

[1] J.L. Gauvain, L. Lamel, G. Adda, ``Partitioning and Transcription of Broadcast News Data," ICSLP'98, 5, pp. 1335-1338, Sydney, Australia, December 1998.

[2] J.L. Gauvain, L. Lamel, G. Adda, ``The LIMSI Broadcast News Transcription System" Speech Communication, 37(1-2):89-108, May 2002.

[3] L. Mangu, E. Brill, A. Stolke, “Finding Consensus Among Words: Lattice-Based Word Error Minimization“, Eurospeeech'99, 495-498, Budapest, Sep. 1999.

[4] J.L. Gauvain and C.H. Lee, “Maximum A Posteriori for Multivariate Gaussian Mixture Observation of Markov Chains“, IEEE Trans. on Speech and Audio Processing, pp. 291-298, 1994.

[5] P.C. Woodland, T. Neieler, E. Whittaker, “Language Modeling in the HTK Hub5 LVCSR“, presented at the 1998 Hub5E Workshop, Sep. 1998.

7.3 Spanish BN Recognizer by RWTH

Spanish TC-STAR
TASK 1 (HUB4 SPANISH BASELINE) - TASK 2 (TCSTAR-P) (- TASK 3 (EPPS))
RWTH

0) INTRODUCTION

The Spanish baseline system for the TCSTAR project is entirely new. That is, the HUB4SP as well as the TCSTAR-P recognizer were trained from scratch within TC-STAR. The general training setup followed the HUB4EN recognizer as described in [2], and resulted in an across-word recognizer using a single, gender independent acoustic model. For the baseline system a single pass was performed, neither MLLR nor VTN was applied. For task 1 the system was run on the HUB4SP 97 evaluation set, available at LDC (LDC2001S91). Task 2 was performed on the TCSTAR-P Spanish evaluation set. Task 3 will consist of testing both baseline systems on one hour parliamentarian speech once the transcribed data will be available within TC-STAR.

1) ACOUSTIC ANALYSIS

We used standard MFCC features.
The magnitude spectrum was estimated by applying the DFT to the preemphasised and windowed audio signal each 10ms. Next the magnitude spectrum was filtered with a filter bank consisting of 20 triangular filters positioned at equidistant points on the Mel frequency axis. The logarithms of the filter outputs were cepstrally decorrelated (discrete cosine transform), resulting in 16 dimensional vectors. The MFCCs were normalised using cepstral mean removal, and energy and variance normalisation. Nine temporally consecutive vectors were fed into an LDA to obtain 45 dimensional feature vectors which were used for the baseline results.

2) ACOUSTIC MODEL

The words of the vocabulary were modeled by position-dependent triphones with across-word contexts [2]. The triphones were represented by Hidden Markov Models (HMMs). The non-silence HMMs used a standard three states left-to-right topology, where each of the states was duplicated resulting in a six states HMM model, whereas the silence HMM consists of a single HMM state. The emission probabilities assigned to the HMM states in turn were modeled by Gaussian mixture models, sharing a single, globally pooled diagonal covariance matrix. The transition probabilities were empirically estimated. HMM states were tied using a binary decision tree (CART). During training and recognition we used the Viterbi approximation on the state-level.

For task 1 the tied states were trained on the HUB4SP 1997 training corpus (LDC98S74/LDC98T29) containing about 30 hours of speech. The lexicon used was an automatically augmented version of the CALLHOME Spanish lexicon, almost identical

to the one used for recognition (see 5.). The acoustic model consists of 2,501 tied states and 270,305 densities. For task 2, the TCSTAR-P Spanish training set (about 7.5 hours of speech) was added to the HUB4SP training corpus. The lexicon was given by the union of the HUB4SP lexicon and the lexicon provided with the TCSTAR-P transcripts. The phoneme sets of both parts of the training data were kept separate, again CART was used to tie HMM states. The set of questions allowed tied states within and between both sets of phonemes. The resulting acoustic model consists of 2,501 generalized phoneme models and 283,925 densities, shared over both phoneme sets.

3) LANGUAGE MODEL

For each task a single n-gram language model was used. The models were estimated by interpolating lower order backoff n-gram-models, where the backoff weights were obtained by absolute discounting (Kneser-Ney). For estimation the SRI language modeling toolkit was used [4]. The models were trained on the following sources, all available via LDC or being part of the TCSTAR-P data. For task 1 we used the 'Spanish News Text', vol. 1, corpus (LDC95T9) and the HUB4SP transcripts (LDC98T29). For task 2 the 'Spanish Newswire Text', vol. 2, corpus (LDC99T41) and the TCSTAR-P transcripts were added. The oov-rates on task 1 and task 2 are 2.1% and 1,2%, respectively. The first number was achieved on the HUB4SP 97 development data and the latter on the TCSTAR-P Spanish development set.

4) RECOGNITION LEXICON

The lexicon used for task 1 is an augmented version of the CALLHOME Spanish lexicon available at LDC (LDC96L16). Words occurring in the test data but not being part of the CALLHOME lexicon were automatically transcribed and added to the lexicon yielding an oov-rate of approximately zero on the training data. The transcription was done by a grapheme-to-phoneme model trained on the CALLHOME lexicon [3]. A phoneme set of size 30 was determined by the CALLHOME lexicon. In addition a silence phoneme, a single phoneme describing filled pauses, and four phonemes describing different kind of noises were added. Finally, the lexicon consists of 36 phonemes and 50,804 words.

For task 2 the lexicon delivered with the TCSTAR-P transcripts was used. In addition to the 32 given phonemes a silence, a filled pause, and a noise phoneme were added. The lexicon contains 12,470 words.

For both lexicons noise events were mapped to silence during recognition.

5) RECOGNITION

Our baseline system was a gender independent, single pass across-word recognizer. A beam search strategy with a pre-pruning step based on language model look-ahead using

a bigram model [1] was applied. Neither VTN nor MLLR were used to produce the baseline results.

6) EXECUTION TIME

On an AMD Athlon MP with 1800Mhz and 3GB RAM a real-time factor of about 11 was measured for task 1. For task 2 the real-time factor was 12 on an AMD Athlon MP with 2000Mhz and 3GB RAM.

7) REFERENCES

[1] A. Sixtus, S. Molau, S. Kanthak, R. Schlüter, H. Ney. “Recent Improvements of the RWTH Large Vocabulary Speech Recognition System on Spontaneous Speech“. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1671-1674, Istanbul, Turkey, June 2000.

[2] A. Sixtus. “Across-Word Phoneme Models for Large Vocabulary Continuous Speech Recognition“. Dissertation, Aachen, Germany, January 2003.

[3] M. Bisani, H. Ney. “Multigram-based Grapheme-to-Phoneme Conversion for LVCSR“. In Proc. European Conference on Speech Communication and Technology, Vol 2, pp. 933-936, Geneva, Switzerland, September 2003.

[4] A. Stolcke. “SRILM - An Extensible Language Modeling Toolkit“. In Proc. International Conference on Spoken Language Processing, Denver, CO, Sept. 2002.

8 System Descriptions Mandarin

8.1 Mandarin BN Recognizer by UKA

1. Primary Test System Description

The ISL RT-04f Mandarin Broadcast News evaluation system uses the JANUS speech recognition toolkit.

The front-end processing is based on 13 MFCC features using a context window of 15 frames. Cepstral mean and variance compensation for each cluster was followed by an LDA transform, giving the final feature vector of 42 components. Vocal tract length normalization was performed on a cluster basis.

Two sets of gender-independent acoustic models were applied: one using an initial-final (IF) lexicon and another using a phone-level lexicon. The IF system has 3000 clustered triphone states and a total of 168k Gaussians; the phone system has 3000 tied septaphone states with a total of 169k Gaussians. Tonal information was used in decision trees such that a single tree was used for all tonal variants of the same phone.

Maximum likelihood training was used for both sets of models. The mixtures were grown incrementally over several iterations. A single global semi-tied covariances (STC) are employed. The acoustic models were trained in a cluster adaptive way, making use of cluster based feature space transforms (FSA-SAT). Speaker adaptation during testing was carried out on the features (FSA), means (MLLR).

The partition strategy consists of four components: speech/non-speech segmentation, music detection, foreign language detection, and speaker clustering. It proceeds in the following steps:

1) initial segmentation using energy-based speech/non-speech detection (CMU segmenter, CMUseg_0.5 package);
2) Gaussian mixture model based music/non-music classification: Music segments were subsequently discarded;
3) Language identification: We use a phonetic language modeling approach to detect English segments in a Chinese show. An open-loop Chinese phone recognizer is used to decode both Chinese BN shows and English BN shows. The output phone sequence is used to train an n-gram phonetic language model, one for Chinese and one for English. During testing, each speech segment is first decoded by the Chinese phone recognizer. Then, the output phone sequence is compared to both the Chinese phonetic language model and the English phonetic language model. The likelihood ratio is used to determine the language identity of the segment. The Chinese phonetic language model is trained on a 2-hour subset of the 1997 Hub4 Mandarin training data. The English phonetic language model is trained on a 5-hour subset of the 1996 BN English training data. Bigram phonetic language model is used in both cases.

4) speaker clustering: All speech segments are clustered using a hierarchical, agglomerative clustering algorithm, which employs a tied-GMM based distance measure and a BIC based stopping criteria.

The Ibis single pass decoder was used to decode the evaluation data. Cross-adaptation between the two sets of acoustic models was performed to progressively refine the hypotheses. A 4-gram language model was further used to rescore lattices from earlier stages. We then applied confusion network combination.

2. Training

The acoustic models were trained on:

a. 27 hours of manually transcribed Broadcast News data released by LDC (LDC98S73)
b. 69 hours of quickly transcribed TDT4 Mandarin data (LDC2003E21)

The language models were trained on:

- Mandarin Chinese News Text Corpus
- China Radio 1994-1996
- People's Daily 1991-1996
- Xinhua News 1994-1996
- TDT2 and TDT3
- TDT4 (excluding text data preceding the last test epoch (Feb 2001))
- Mandarin Gigaword corpus
- Xinhua News 1990 - 2002 (excluding text data preceding the last test epoch (Feb 2001))
- HUB4m 1997 training transcript
- RFA (web-crawled) from 2001 (excluding text data preceding the last test epoch (Feb 2001)) to Nov 2003
- NTDTV (web-crawled) from 2002 to Nov 2003

We incorporated the LDC name entity list into our text segmenter's wordlist and then segmented the text data. Then we derived the word vocabulary from the segmented text. We added the Chinese character set of size 6.7k to the vocabulary. The size of the vocabulary is around 63k. We employed count-mixing approach to train the word trigram and 4-gram LMs. The mixing weight for HUB4m 1997 transcript is set to 6 while the mixing weight for other text sources are set to 1. We used the SRI LM toolkit to train the LM. The LMs were smoothed using Kneser-Ney smoothing scheme. We pruned word trigram and word 4-gram counts by applying count cutoff. The minimum counts of word trigram and 4-gram are 3 and 5 respectively.

The lexicon contains 84K entries derived from the LDC CallHome Mandarin lexicon (LDC96L15). We used a maximal matching technique to generate pronunciations for words not in the LDC lexicon. There are 23 Initials and 34 Finals in the initial-final model, and 38 phonemes in the phone-based models. Eight additional phonemes are used for noises and silence.

3. Execution Time

Processing of the test data took about 26 times real-time on a 3.2G Pentium4 single CPU Linux box. Process size was about 600MB during decoding.

4. References

[1] P. Zhan, S. Wegmann and S. Lowe.
Dragon Systems' 1997 Mandarin Broadcast News System.
DARPA Broadcast News Workshop, 1999

[2] L. Nguyen, B. Xiang and D. Xu.
The BBN RT03 BN Mandarin System.
RT-03 Workshop, Boston, 2003

[3] D. Liu, J. Ma, D. Xu, A. Srivastava, F. Kubala.
Real-Time Rich-Content Transcription of Chinese Broadcast News.
ICSLP, Denver, 2002

[4] Hagen Soltau, Florian Metze, Christian Fuegen, and Alex Waibel.
A One-pass decoder based on polymorphic linguistic context assignment.
ASRU 2001

[5] Mark Gales.
Semi-Tied Full-Covariance Matrices for Hidden Markov Models.
Technical Report, Cambridge University 1997

[6] Mark Gales.
Maximum Likelihood Linear Transformations for HMM-based Speech Recognition. Technical Report, Cambridge University 1997

[7] Lidia Mangu, Erik Brill, and Andreas Stolcke.
Finding Consensus among words: Lattice-based word error minimization.
Eurospeech 99

[8] Hua Yu and Alex Waibel.
Streamlining the Front-End of a Speech Recognizer.
ICSLP, Beijing, 2000

[9] H. Soltau, H. Yu, F. Metze, C. Fuegen, Q. Jin and S. Jou.
The ISL Transcription System for Conversational Telephony Speech.
ICASSP, Montreal, 2004

[10] Q. Jin, K. Laskowski, T. Schultz, and A. Waibel.
Speaker Segmentation and Clustering in Meetings.
NIST Meeting Recognition Workshop, Montreal, May 2004.

[11] H. Yu and T. Schultz.
Enhanced Tree Clustering with Single Pronunciation Dictionary for Conversational Speech Recognition.
Eurospeech, Geneva, 2003.

[12] P. Delacourt and C.J. Wellekens.
DISTBIC: A Speaker-based Segmentation for Audio Data Indexing.
Speech Communications, 32, 111-126, 2000.

[13] S. Chen and P.S. Gopalakrishnan.

Speaker, Environment and Channel Change Detection and Clustering via Bayesian Information Criterion.
DARPA Speech Recognition Workshop, 1998.

[14] Marc A. Zissman.
Language Identification Using Phone Recognition and Phonotactic Language Modeling.
ICASSP, Volume 5, pp 3503-3506, Detroit, May 1995.

8.2 Mandarin BN Recognizer by LIMSI

LIMSI Mandarin baseline system

The LIMSI Mandarin Broadcast News system is essentially the same as that used in the DARPA RT03 HUB-4NE 10x evaluation [8], with models (lexicon, acoustic models, language models) trained for Mandarin Chinese.

1) GENERAL SYSTEM DESCRIPTION:

The LIMSI segmentation and clustering is based on an audio stream mixture model [1,2]. First, the non-speech segments are detected and rejected using GMMs representing speech, speech over music, noisy speech, pure-music and other background conditions. An iterative maximum likelihood segmentation/clustering procedure is then applied to the speech segments. The result of the procedure is a sequence of non-overlapping segments with their associated segment cluster labels. Each segment cluster is assumed to represent one speaker in a particular acoustic environment and is modeled by a GMM. The objective function is the GMM log-likelihood penalized by the number of segments and the number of clusters, appropriately weighted. Four sets of GMMs are then used to identify telephone segments and the speaker gender. Segments longer than 30s are chopped into smaller pieces by locating the most probable pause within 15s to 30s from the previous cut.

The LIMSI BN speech recognizer [2] uses 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band (or 0-3.5kHz for telephone data) every 10ms. For each 30ms frame the Mel scale power spectrum is computed, and the cubic root taken followed by an inverse Fourier transform. Then LPC-based cepstrum coefficients are computed. The cepstral coefficients are normalized on a segment-cluster basis using cepstral mean removal and variance normalization. Thus each cepstral coefficient for each cluster has a zero mean and unity variance. The 39-component acoustic feature vector consists of 12 cepstrum coefficients and the log energy, along with the first and second order derivatives.

Each phone model is a tied-state left-to-right CD-HMM with Gaussian mixtures. The triphone-based context-dependent phone models are word-independent but position-dependent. The tied states are obtained by means of a decision tree.

Word recognition is performed in three passes, where each decoding pass generates a word lattice which is expanded with a 4-gram LM. The posterior probabilities of the lattice edges are estimated using the forward-backward algorithm. The 4-gram lattices are converted to a confusion network with posterior probabilities by iteratively merging lattice vertices and splitting lattices edges until a linear graph is obtained. This procedure gives comparable results to the edge clustering algorithm proposed in [3]. The words with the highest posterior in each confusion set are hypothesized.

Pass 1: Initial Hypothesis Generation - This step generates initial hypotheses which are then used for cluster-based acoustic model adaptation. This is done via one pass (less than 1xRT) cross-word trigram decoding with gender-specific sets of position-dependent triphones (5500 tied states) and a trigram language model (8M trigrams and 8M bigrams). Band-limited acoustic models are used for the telephone speech segments. The trigram lattices are rescored with a 4-gram language models.

Pass 2: Word Graph Generation - Unsupervised acoustic model adaptation is performed for each segment cluster using the MLLR technique [4] with only one regression class. The lattice is generated for each segment using a bigram LM and position-dependent triphones with 11500 tied states (32 Gaussians per state).

Pass 3: Word Graph rescoring - The word graph generated in pass 2 is rescored after carrying out unsupervised MLLR acoustic model adaptation using two regression classes.

2) ACOUSTIC TRAINING:

The acoustic models were trained on about 27 hours of Hub4-Mandarin training data (from LDC) and about 100 hours of data from the TDT4 corpus. Since time-aligned transcripts are not available, the TDT4 data from the Mainland China sources (CNR, CTV and VOA) were transcribed with our recognizer using acoustic models estimated on the manually transcribed Hub4-Mandarin data and with source-specific language models estimated on the TDT4 closed captions for each source. Wide-band and bandlimited models were trained by pooling the Hub4 Mandarin data and the TDT4 data from Mainland China.

The acoustic models are position-dependent triphones with tied states, obtained using a divisive decision tree based clustering algorithm. Two sets of gender-dependent acoustic models were built using both MAP adaptation [5] of SI seed models for each of wideband and telephone band speech.

The English Hub4 training data was used to build the Gaussian mixture models for gender identification, and music and telephone segment detection. About 2 hours of pure music portions of the acoustic training data were used to estimate the music GMM.

3) LANGUAGE MODEL TRAINING

The n-gram language models were obtained by interpolation [6] of backoff n-gram language models trained on the following sources available via LDC:

1 - TDT2,3,4 Mandarin transcripts		(10.2M characters)
2 - People Daily newspaper	1991-1996	(85M characters)
3 - China Radio transcripts	1994-1996	(87M characters)
4 - Xinhua news	1994-1996	(22M characters)

as well as additional data shared with us by BBN from Mainland

5- People Daily newspaper 1997,1999,2000 (39M characters)

Different component LMs are trained on the text sources mentioned above, with the mixture weights optimized using the transcriptions of dev03 data. The interpolation coefficients were chosen in order to minimize the perplexity a set of dev03 shows (and transcripts) shared by BBN. The dev03 shows are:

20001226_2000_2025_CTS_MAN
20001226_1700_1730_CNR_MAN
20001227_0800_0820_CBS_MAN
20001229_1330_1400_CTV_MAN
20001231_0700_0800_VOA_MAN

The 57k word list was selected from the same text sources so as to minimize the OOV rate on the dev03 data. The word list contains 57703 entries, including all characters (i.e., there are essentially no OOV characters).

4) RECOGNITION LEXICON DESCRIPTION

Pronunciations are based on a 61 phone set (4 of them are used for silence, filler words, and breath noises). The 5 tones for the vowels are collapsed into 3 tones for each vowel (rising, flat and falling) A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones. The 57k vocabulary contains 57707 words with 59152 phone transcriptions.

5) REFERENCES

[1] J.L. Gauvain, L. Lamel, G. Adda, ``Partitioning and Transcription of Broadcast News Data," ICSLP'98, 5, pp. 1335-1338, Sydney, Australia, December 1998.

[2] J.L. Gauvain, L. Lamel, G. Adda, ``The LIMSI Broadcast News Transcription System" Speech Communication, 37(1-2):89-108, May 2002.

[3] L. Mangu, E. Brill, A. Stolke, “Finding Consensus Among Words: Lattice-Based Word Error Minimization“, Eurospeeech'99, 495-498, Budapest, Sep. 1999.

[4] C.J. Legetter, P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,“ Computer Speech & Language, Vol. 9, pp. 171-185, 1995.

[5] J.L. Gauvain and C.H. Lee, “Maximum A Posteriori for Multivariate Gaussian Mixture Observation of Markov Chains“, IEEE Trans. on Speech and Audio Processing, pp. 291-298, 1994.

[6] P.C. Woodland, T. Neieler, E. Whittaker, “Language Modeling in the HTK Hub5 LVCSR“, presented at the 1998 Hub5E Workshop, Sep. 1998.

[7] L. Chen, J.L. Gauvain, L. Lamel, and G. Adda, “Unsupervised Language Model Adaptation for Broadcast News“, ICASSP'03.

[8] L. Chen, L. Lamel, G. Adda, and J.L. Gauvain. “Broadcast News Transcription in Mandarin.” In Proc. ICSLP'2000, pages II-1015-1018, Beijing, Oct 2000.