The LIMSI 1995 Hub3 System

J.L. Gauvain, L. Lamel, G. Adda, D. Matrouf

LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{gauvain,lamel,gadda,driss}@limsi.fr

ABSTRACT

In this paper we report on the LIMSI recognizer evaluated in the ARPA 1995 North American Business (NAB) News Hub 3 benchmark test. The LIMSI recognizer is an HMM-based system with Gaussian mixture. Decoding is carried out in multiple forward acoustic passes, where more refined acoustic and language models are used in successive passes and information is transmitted via word graphs. In order to deal with the varied acoustic conditions, channel compensation is performed iteratively, refining the noise estimates before the first three decoding passes. The final decoding pass is carried out with speaker-adapted models obtained via unsupervised adaptation using the MLLR method. In contrast to previous evaluations, the new Hub 3 test aimed at improving basic SI, CSR performance on unlimited-vocabulary read speech recorded under more varied acoustical conditions (background environmental noise and unknown microphones). On the Sennheiser microphone (average SNR 29dB) a word error of 9.1% was obtained, which can be compared to 17.5% on the secondary microphone data (average SNR 15dB) using the same recognition system.

INTRODUCTION

In this paper we report on the LIMSI speech recognizer used in the ARPA November 1995 evaluation on the North American Business (NAB) News task[13]. LIMSI has participated in annual ARPA sponsored continuous speech recognition evaluations aimed at improving basic speech recognition technology since November 1992. The goal of the 1995 Hub 3 task was to "improve basic speaker-independent performance on unlimited-vocabulary read speech under acoustical conditions that are somewhat more varied and degraded than speech used in previous ARPA evaluations". Besides the problems posed by the unlimited vocabulary dictation task on reasonably clean speech data (such as the WSJ0/WSJ1 corpus), one of the major challenges of the Nov95 evaluation was to achieve acceptable performance on other (ie. non close-talking) microphone data with no prior knowledge of either the microphone type or the background noise characteristics.

In the next section we provide an overview of the LIMSI speech recognition system and the decoder strategy. We then describe our development work in language modeling, including the text processing and vocabulary selection. The recognition lexicon is presented along with a description of our semi-automatic method for adding pronunciations for new words. We then return to the experiments carried out with acoustic modeling and environmental compensation aimed at improving performance on the noisy data. In contrast to previous evaluations, where for the primary system each sentence was treated independently (i.e., the results must be independent of the order in which the test sentences were processed), this year we used the knowledge of the article boundaries and utterance order to carry out unsupervised transcription-mode adaptation.

RECOGNIZER OVERVIEW

The LIMSI speech recognizer makes use of continuous density HMMs with Gaussian mixture for acoustic modeling and n-gram statistics estimated on newspaper texts for language modeling. The recognition vocabulary contains 65k words selected to minimize the out-of-vocabulary rate on a set-aside portion of the development text set. Bigram and trigram language models were trained on 284M words of text and read WSJ0/1 speech transcriptions predating *July 30, 1995* (inclusive). Context-dependent phone models were trained on the Sennheiser channel on 46k sentences taken from the WSJ0/1 corpus. The decoding is carried out in multiple passes, with more accurate models in successive passes. All passes use cross-word CD phone models. Revised noise estimates are made in between decoding passes and unsupervised speaker adaptation is carried out in the final pass.

Acoustic models

Acoustic modeling uses 48 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band every 10ms (30ms window). Cepstral mean removal was performed for each sentence. The models were trained on 46,146 sentences (about 99 hours of speech) from 355 speakers of the WSJ0/1 corpus. This is comprised of 37,518 sentences from the WSJ0/1 SI-284 corpus, 130 sentences/speaker from 57 long-term and journalist speakers in WSJ0/1, and 1218 sentences from 14 of the 17 additional WSJ0 speakers not included in SI-84. Only the data from the close-talking Sennheiser HMD-410 microphone was used for training.

Each phone model is a tied-state left-to-right, 3-state CDHMM with Gaussian mixture observation densities (typically 32 components). The triphone contexts to be modeled are selected based on their frequencies in the training data, with backoff to right-context, left-context, and context-independent phone models.

Separate male and female models obtained with MAP estimation[5] are used to more accurately model the speech data. Different size models were built for use in successive decoding passes. The model sets used in this evaluation were:

- two sets of speaker-dependent 490 CD phone models, 31 Gaussians per state (total of 45k Gaussians)

- two sets of 3500 gender-dependent CD phone models with 6000 tied states, 31 Gaussians per state (190k Gaussians per model set);

- two sets of 5300 gender-dependent CD phone models with 7000 tied states, and 31 Gaussians per state (total of 220k Gaussians per model set);

- two sets of 7900 gender-dependent CD phone models with 10400 tied states, and 31 Gaussians per state (total of 325k Gaussians per model set);

The smallest model set was used only for gender selection and endpoint detection. The middle two model sets were used for adaptation, and the largest model set was used only for a contrast condition with the clean-speech test data (C0).

Decoding

Decoding is carried out in multiple passes, where more accurate acoustic and language models are used in successive passes.

- **Step 0: Gender-identification and endpoint detection** Gender identification is performed by running a phone recognizer on *all the data from the given test speaker* and selecting the gender associated with the model set giving the highest likelihood on the entire set[9]. Gender identification uses a small acoustic model set (490 SI, CD models) with a phone bigram to provide phonotactic constraints. The sentence initial and final silences are removed in this pass.

- **Step 1: Bigram decoding** A word graph is generated using a bigram LM. Due to memory constraints, this step is actually carried out in two passes, the first with gender-specific sets of 3500 *position-dependent* triphone models and a small bigram LM (cutoff 10) and the second with gender-specific sets of 5300 *position-independent* context-dependent phone models and a larger bigram LM (cutoff 1).

- **Step 2: Trigram decoding** The sentence is decoded using the same set of 5300 gender-specific *position-independent* phone models and the word graph generated by the 2nd bigram pass, with the trigram language model. This step is also carried out in 2 passes. The first pass uses a more compact trigram LM (cutoffs 1 and 2), and the second pass uses a larger trigram LM (cutoffs 0 and 1) with speaker-adapted models obtained via unsupervised adaptation using the MLLR method[10]).

Compared to the LIMSI recognizer described previously[6, 7, 8], this year's system has the following new attributes:

- State-tying is used to reduce the size of the acoustic models in order to facilitate model adaptation (for noise compensation and speaker adaptation) and to increase the triphone coverage of a larger set of clean speech models;

- Noise compensation is performed for additive and convolutional noises (to facilitate this, the log energy has been replaced by the first cepstral coefficient);

- Gender selection is based on all the data from a given speaker, rather than on a sentence-by-sentence basis;

- *Position-dependent* triphones are used in the first decoding pass so as to optimize the coverage of the cross word triphones versus the number of models (given memory limitations);

- Unsupervised speaker adaptation using the MLLR method is used to create speaker-specific acoustic models for the final decoding pass.

MODEL ADAPTATION

Since no prior knowledge of either the microphone type, the background noise characteristics or the speaker identity is available to the system, model adaptation has to be peformed by using only the data in the test, i.e. in unsupervised mode.

Environmental adaptation is based on the following model of the observed signal y given the input signal x: $y = (x + n) * h$, where n is the additive noise and h the convolutional noise. Compensation is performed iteratively, where refined estimates of n and h are obtained before each of the first three passes of the decoding process (gender identification and the two bigram passes). Estimation makes use of the 3s background sample provided for each speaker session, the silence segments from the test material (not used in the first phone recognition pass) and a Gaussian model of the test speech (the 15 test sentences). The compensated models are obtained by adapting models trained exclusively on the Sennheiser data. We use a data driven approach which is related to model combination schemes[3, 11, 4].

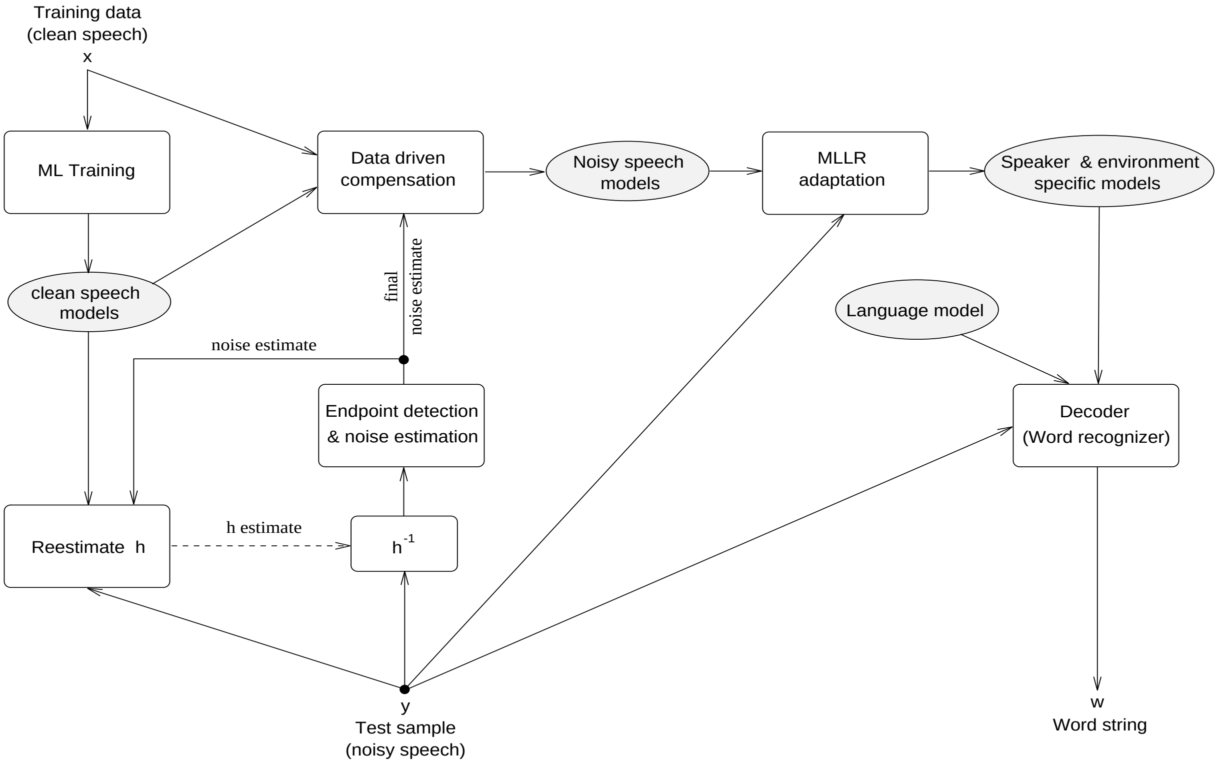

Figure 1: Model adaptation process.

Parallel model combination (PMC) approximates a noisy speech model by combining a clean speech model with a noise model. For practical reasons, it is generally assumed that the noise density is Gaussian and that the noisy speech model has the same structure and number of parameters as the clean speech model – typically a continuous density HMM with Gaussian mixture. Various techniques have been proposed to estimate the noisy speech models, including the log-normal approximation approach, the numerical integration approach, and the data driven approach[4]. The log-normal approximation is crude especially for the derivative parameters, and all three approaches require making some approximations to estimate non-trivial derivative parameters.

For this work we have chosen to use a data-driven approach, where in order to avoid making all the approximations of model combination, we directly use the original clean speech training samples instead of generating clean speech samples from the clean speech models. In order to be efficient, the approach requires (like data-driven PMC) the precomputation and clipping of the Gaussian posterior probabilities for a given training frame. These values are assumed to remain unchanged after adding the noise frames to the clean speech frames. In comparison to other proposed approaches, this scheme is computationally inexpensive, but requires reading all of the clean speech training data from disk. However, with proper organisation and compression of the training data, we have observed that model adaptation using this scheme can in fact be performed faster than by using PMC with the log-normal approximation approach. This is true even with relatively large amounts of training data (on the order of 20h of speech) since more parameters are typically used when more training data is available. (With the log normal approach the processing time is roughly proportional to the model size, where as with the data-driven approach it is proportional to the training data size.)

In addition to allowing the use of any kind of derivative parameters, the data-driven approach also allows the use of sentence-based cepstral mean removal, which is commonly used to make the acoustic features robust to convolutional noise. However, this can only be done properly if the additive noise n can be estimated from the observed noise $h * n$, or equivalently, if the convolutional noise h can be estimated for the noisy speech sample. The noise n can be estimated iteratively starting with the silence frames n_0 of the adaptation data (noisy test data). These silence frames are used to compute the noisy speech cepstrum mean (using log-normal approximation PMC or data-driven PMC), which is substracted from the cepstrum mean of adaptation data to obtain a first estimate of $\tilde{h}$. The filter $\tilde{h}^{-1}$ is then applied to the adaptation data to obtain a better estimate of n. We observed that in practice no more than 5 iterations are needed to properly estimate n and h. (It should be noted that cepstral mean removal is not performed when estimating h.) The model adaptation process is shown in Figure 1.

Unsupervised speaker adaptation performed in the last decoding pass is based on the ML linear regression technique[10]. A single full regression matrix (49 × 48) is used to transform the Gaussian means of the models for the hypothesized gender. The use of a single regression matrix makes speaker adaptation effective even with the high recognition error rates on the low SNR data.

LANGUAGE MODELS AND LEXICON

In this section we describe our development work for language modeling and modifications to our recognition lexicon.

Selection of a development text set

Development texts were needed to choose the recognition vocabulary so as to maximize coverage and to measure the perplexity of different language models we estimated. Since the dev95 speech data used the same prompt text as were used for the dev94 speech data, and no separate set of development texts were distributed for 1995, we selected a subset of the Hub3-95 texts according to the NIST criteria (source distribution, articles and sentences selection and verification). A set of 600 sentences (17k words) were selected from the texts of the last day of July 1995 (about 200,000 words) so as to have no overlap between the development, the training and the evaluation texts. This also represented a compromise in terms of size: enough texts to select a subset from, but not too large of a set to remove from the training data.

We actually selected two subsets of development texts, with and without headlines (eg. `Beirut LEBANON Reuters`), but we observed no difference with respect to out-of-vocabulary rates (OOV) and perplexity measures.

Text preprocessing

We preprocessed all the 1995 Hub3 and Hub4 training texts, predating the 31st of July, as well as then August'94 release of the CSR standard LM training texts. Starting with the standard verbalized punctuation form distributed by LDC, training texts were cleaned to remove errors inherent to the texts or arising from processing with the distributed text processing tools.

As was done last year, the texts were transformed to be closer to the observed American English reading style[7, 8]. The set of rules and the corresponding probabilities were derived from the examination of the WSJ1/WSJ0 acoustic data (prompts and transcriptions). For example, while the default text processing tools convert $1/8$ into *one eighth*, people say *an eighth* just as frequently, so a rules maps 50% of the former into the latter.

Word list selection

To select the 65k word list, OOV rates on three text sets were compared: last year's development texts (*dev95*), last year's evaluation texts (*eval94*), and the LIMSI 1995 development text set (*l-dev95*), using different subsets of the available text data. Different combinations of the texts were tried, and the 65,500 most common words selected. The OOV rates on the three test sets are shown in Table 1. We observed that combinations which included the Hub4 (H4) data typically had higher OOV rates than without this data. The chosen combination was: *WSJ92-94* (45M words), Hub3 (*H3*, 44M words), *dev94* (1.9M words), and *WSJ0/WSJ1* read speech transcriptions (consisting of 85k sentences and 1.4M words). Weighting the *dev94* texts and the transcriptions by 2 gave the lowest OOV rate on the development data and minimized the number of new words to be added to the lexicon.

Text Source	*dev94*	*eval94*	*l-dev95*
87-94	49	56	138
H3	33	35	95
H3+H4	37	44	126
WSJ	40	35	119
WSJ+dev94+si85k	31	35	119
WSJ+dev94+si85k+H3	25	33	92
WSJ+dev94+si85k+H3+H4	33	38	104

Table 1: # OOVs using different text sources:
87-94 : all standard training material from 87 to July 94
WSJ : WSJ subset from **92-94**
H3 : all texts from the Hub3 LM material (< 31st of July)
H4 : all texts from the Hub4 LM material (< 31st of July)
dev94 : 1994 NAB development data (excluding *dev94* set)
si85k : WSJ0/WSJ1 read speech transcriptions

Text	*CMU 60k*		*LIMSI 65k*		*OOV rate*
set	*#*	*%*	*#*	*%*	*reduction*
dev94	44	(0.6)	26	(0.4)	40%
eval94	59	(0.7)	31	(0.4)	48%
l-dev95	143	(0.9)	92	(0.6)	36%
eval95	60	(1.0)	47	(0.8)	20%

Table 2: Comparison of OOV rates for the standard CMU 60k and the LIMSI 65k word lists.

Table 2 compares the OOV rates of the LIMSI Nov95 65k vocabulary and the standard CMU 60k vocabulary. While on the different dev sets the LIMSI OOV rate is about 40% lower than the CMU OOV rate, on the 1995 eval data, only a 20% OOV rate reduction was obtained. Apparently the development set *l-dev95* was not a very good estimator of the evaluation texts.

Language models

We estimated language models on all the Hub3 available training texts (excluding the texts of July 31, 1995), the read speech transcriptions, and optionally, the Hub4 language model training texts. Four language models were constructed using the CMU toolkit, ranging from a small bigram in the first pass to a large trigram for the final pass.

Language model perplexities with and without the Hub4 texts are given in Table 3. With a small bigram the perplexity increased on all 3 development text sets. For the large trigram model, small but inconsistent differences were observed, so

Training text	87-94+H3	87-94+H3+H4
cutoff	10	14
# bg (M)	1.56	1.56
dev94	236.8	242.4
eval94	249.1	254.8
l-dev95	234.1	241.7
cutoffs	0-1	0-1
# bg-tg (M)	15.7-21.1	19.7-28.9
dev94	130.2	128.8
eval94	136.3	133.9
l-dev95	126.0	126.5

Table 3: Perplexity of the development text sets, for bigram and trigram trained with and without Hub4 texts.

we decided not to use the Hub4 training texts.

The training texts were reprocessed in order to obtain a second version in which the 1000 most frequent acronyms are treated as whole words instead of as sequences of independent letters. The motivation was two-fold: to have better word level context modeling in the language model, and to more easily represent reduced pronunciation variants for common acronyms such as S&L, AT&T, IRA, IRS, where the middle word is often highly reduced.

A new 65k word list and bigram and trigram LMs were built with acronyms. In order to compare the perplexity to that obtained with the original LMs we needed to normalize for the difference in text length n (about 1%). We used a normalized perplexity p^* defined as:

$$p^* = 2^{\frac{n_1}{n_2}\frac{log(p)}{log(2)}}$$

where n_1 and n_2 are the number of words in the text with and without acronyms respectively, and p and p^* are the unnormalized and normalized perplexity on the text with acronyms.

Using the normalized perplexity, the bigram perplexity of the text with acronyms is actually about 5% lower than for the original texts. The bigram and trigram perplexities on the transcriptions of the ARPA Nov95 evaluation data are 239.3 and 137.2, respectively.

Conditions	*dev94*	*eval94*	*l-dev95*
no acronyms, px	236.8	249.1	234.1
acronyms, px	245.9	256.8	245.8
acronyms, normalized px	231.9	244.7	225.7

Table 4: Perplexity and normalized perplexity, on different developement text sets, for training texts with and without acronyms.

Recognition lexicon

Creation of pronunciation lexicons for speech recognition is widely acknowledged to be an important aspect of system development, that is labor-intensive. Lexicons are often manually created and make use of knowledge and expertise that is difficult to codify. Our experience in large vocabulary, continuous speech recognition is that systematic lexical design can improve the overall system performance. Our approach is to represent the lexicons with standard pronunciations using a set of 45 phonemes and do not explicitly represent allophones. We have chosen a phonemic representation, as most allophonic variants can be predicted by rules, and their use is optional. More importantly, there often is a continuum between different allophones of a given phoneme and the decision as to which occured in any given utterance is subjective. By using a phonemic representation, no hard decision is imposed, and the acoustic models can automatically learn the observed variants in the training data. Frequent alternative variants which are not allophonic differences (such as the suffix *"-ization"* in American English which can be pronounced with a diphthong (/Y/) or a schwa (/x/)) are explicitly represented in the lexicon. These frequent inflected forms have been verified to provide more systematic pronunciations.

Since generating pronunciations is time-consuming and error-prone (it is mostly manual work), several utilities were developed to facilitate the work. While these utilities can be run in an automatic mode, our experience that human verification is required, and that interactive use is more efficient.[1] First, missing pronunciations are generated by rule when possible, by automatically adding and removing affixes.[2] When multiple pronunciations can be derived they are presented for selection, along with their source. The source lexicons that we make use of are (in order of decreasing confidence): the LIMSI "Master" lexicon, which contains pronunciations for 80k words; the TIMIT lexicon; a modified version of the Moby pronouncing lexicon; and a modified version of the Merriam Webster Pocket dictionary. The Carnegie Mellon Pronouncing Dictionary (version cmudict.0.3) and the Merriam Webster American English Pronouncing Dictionary (book) are also used for reference. We observed that often when no rules applied, it was because the missing word was actually a compound word, or an inflected form of a compound word. Thus, the ability to easily split such words and concatenate the result of multiple rule applications was added.

We evaluate the lexicon in the context of our recognizer by confronting the pronunciations with large corpora. By carrying out a forced alignment of the training data using its orthographic transcription, we are able to estimate the relative frequencies of different alternative pronunciations, as well as to determine sources of pronunciation errors. While it is difficult to systematically evaluate the changes to the lexicon, because in the LIMSI system the set of context dependent

[1] An erroneous transcription early on was obtained for the word "used". The program derived the pronunciation /∧st/, from the word "us". These types of errors can only be detected manually.

[2] This algorithm was inspired by a set of rules written by David Shipman while he was at MIT.

INTEREST	IntrIst In{t}XIst
CIVILIZATION	sIvL[xY]zeSxn
EXCUSE	Ekskyu[sz]

Figure 2: Example lexical entries, with phones in {} being optional, phones in [ ] being alternates.

acoustic models changes when the lexicon is changed, we have observed small but consistent improvements (on the order of 5%) across individual test sets.

The 65,500 word lexicon used in the Nov95 evaluation contains 65,500 words and 72,637 phone transcriptions, with an average of 6.5 phones per transcription. A pronunciation graph is associated with each word so as to allow for alternate pronunciations, which occur for about 10% of the entries. Some example entries are shown in Figure 2. The first word "INTEREST", may be produced with 2 or 3 syllables, depending upon the speaker, where in the latter case the /t/ may be deleted. "CIVILIZATION" illustrates the /x,Y/ alternates mentioned above. In contrast, the alternate pronunciations for "EXCUSE" reflect different parts of speech (verb or noun).

EVALUATION RESULTS

In our development work we made use of the data from 10 speakers of the development set collected by NIST and made available to test participants. This multi-microphone (MUM) corpus contains simultaneous recordings on 8 microphone channels for a variety of background noise levels ranging from 47 to 61dBA[1]. However, since the prompt texts corresponding to this data date from June 1994, the new language models cannot be properly applied to this data.

The Nov95 test data consist of 15 sentences from each of 20 speakers (10m/10f), with simultaneous recordings on two different microphone channels per speaker. The primary test condition (P0) makes use of the secondary microphone channel, and the required contrast condition (C0) makes use of the Sennheiser HMD-410 microphone data. The same recognition system is to be used for both P0 and C0. The P0 data sample 3 different microphones, with all the sentences of each speaker derived from the same microphone. The test prompt texts were extracted from the North American Business (NAB) news texts during the 1-31 August 1995.

Table 5 gives the word error rates obtained on the evaluation data for the P0 and C0 data, with different acoustic models (speaker-adapted or not, noise compensation (yes,no,SNR switch)) and different language models (2-gram and 3-gram). The acoustic model sets were trained only on the clean speech data (the Sennheiser microphone) in the WSJ0/1 corpus. Comparing the first and second lines in the Table, we observe a relative error reduction using a trigram LM of 14% on the P0 data and 21% on the C0 data. In the evaluation system, channel compensation was systematically applied, even for the clean data. The word error on the C0 data without compensation (third line in Table 5) is the same, thus noise compensation doesn't increase error rate on the clean speech. Based on partial runs on the development data, we estimate the word error on the P0 data without channel compensation to be at least 50%.[3] The final decoding pass makes use of a larger trigram LM and speaker-adapted models. An error reduction of 15% is obtained on the P0 data and 13% on the C0 data. The gain is slightly larger for the noisy data because the MLLR adaptation also compensates for some residual mismatch not represented in our channel model. However, even with noise compensation the word error is still twice as high as for the clean speech condition.

Grammar	*Noise*	*Speaker*	*% Word Error*	
condition	*compens.*	*adapt.*	*P0 data*	*C0 data*
2-g	yes	no	23.7	13.2
3-g	yes	no	20.5	10.4
3-g	no	no	>50	10.4
3-g	yes	yes	17.5	9.1
3-g	sw	yes	17.5	8.6

Table 5: Word error rates on the ARPA Nov95 test data for different acoustic and language models: P0 and C0 denote respectively the secondary microphone data and the Sennheiser data.

spkrs	*C0 data*		*P0 data*		P0/C0
	SNR	*%werr*	*SNR*	*%werr*	*werr ratio*
7	28.3dB	7.4	16.3dB	11.6	1.57
7	28.8dB	7.6	15.7dB	14.2	1.87
6	29.9dB	13.1	13.2dB	28.7	2.19

Table 6: Average SNR and word error rates on the three subsets of the ARPA Nov95 test data, each subset represents a primary and secondary microphone pairing.

A contrast condition was also carried out where channel compensation was only performed when the SNR was lower than 25dB, allowing us to use larger sets of acoustic models for clean speech (i.e. SNR higher than 25dB). Each set of clean-speech gender-specific models includes 7895 tied-state context-dependent phone models obtained via MAP estimation[5]. The test data SNR was estimated for each speaker by computing the ratio of the average short term RMS powers of the speech samples and noise samples on a 30ms window after preemphasis with a 0.95 coefficient. The speech/noise decision was based on a bimodal distribution estimated by fitting a mixture of 2 Gaussians to the log-RMS power for all frames[2]. With this configuration a word error of 8.6% was obtained on the C0 data (last row of Table 5). No noise compensation was carried out on the high SNR data, all other system parameters were identical.

In Table 6 the relative increase in word error for the P0

[3] The computation time to process the P0 data without noise compensation exceeds our curiosity to have a more accurate estimate of the word error.

data is shown for the 3 subsets of data corresponding to different secondary microphones. The average SNRs (as defined above) and word errors are given for both sets of data. While the largest word error increase is observed for the lowest SNR (set 3), the difference in SNR between sets 1 and 2 is small, but the increase in word error rate is larger for set2. This suggests that factors, such as changes in microphone characteristics and positioning are not properly compensated with our channel model.

CONCLUSION

In this paper we have described the LIMSI recognizer evaluated in the Nov95 ARPA NAB benchmark test, using multi-microphone data recorded in a variety of background noise conditions. New features of this year's system were channel compensation based on a data-driven approach, state tying to reduce the size of the acoustic models in order to facilitate model adaptation, the use of position-dependent triphones for the first pass so as to optimize the coverage of the cross word triphones versus the number of models and unsupervised speaker-adaptation using the MLLR method in a final decoding pass. We also reprocessed the language model training text materials so as to be able to model the most common 1000 acronyms as words, instead of as sequences of independent letters. The language models were trained on all the available training texts, with the exception of texts from July 31st. The lowest out-of-vocabulary rate was obtained using only a subset of the training texts. The word error obtained on the multi-microphone P0 data was 17.5%. Environmental adaptation based on the $y = (x + n) * h$ model was demonstrated to be effective as it reduced the estimated word error from over 50% without compensation to 17.5% with compensation. Using the same system on the Sennheiser C0 data, a word error of 9.1% was obtained. When channel compensation was applied only for low SNR (less than 25dB), we are able to use a larger sets of acoustic models for the high SNR data, and obtained a word error of 8.6% on the C0 data.

REFERENCES

[1] "Multi-Microphone Data Collection System and Procedures," NIST Speech Disc R27-6.1, Oct. 1995.

[2] D. VanCompernolle, "Noise adaptation in a hidden Markov model speech recognition system", *Computer Speech & Language*, **3**(2), 1989.

[3] M. Gales, S. Young, "An improved approach to hidden Markov model decomposition of speech and noise," *ICASSP-92*.

[4] M. Gales, S. Young, "Robust Continuous Speech Recognition using Parallel Model Combination," *Computer Speech & Language*, **9**(4), Oct. 1995.

[5] J.L. Gauvain, C.H. Lee, "Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities," *Speech Communication*, **11**(2-3), 1992.

[6] J.L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, "Speaker-Independent Continuous Speech Dictation," *Speech Communication*, **15**(1-2), Oct. 1994.

[7] J.L. Gauvain, L. Lamel, M. Adda-Decker, "Developments in Large Vocabulary Dictation: The LIMSI Nov94 NAB System," *ARPA SLS Technology Workshop*, Jan. 1995.

[8] J.L. Gauvain, L. Lamel, M. Adda-Decker, "Developments in Continuous Speech Dictation using the ARPA WSJ Task," *ICASSP-95*.

[9] L. Lamel, J.L. Gauvain, "A Phone-based Approach to Non-Linguistic Speech Feature Identification," *Computer Speech & Language*, **9**, 1995.

[10] C. Legetter, P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, **9**(2), 1995.

[11] F. Martin, K. Shikano, Y. Minami, "Recognition of Noisy Speech by Composition of Hidden Markov Models," *EuroSpeech'93*.

[12] D. Pallett et al., "1994 Benchmark Tests for the ARPA Spoken Language Program," *ARPA SLS Technology Workshop*, Jan. 1995.

[13] J.L. Gauvain, L.F. Lamel, G. Adda, D. Matrouf, "Developments in Continuous Speech Dictation using the 1995 ARPA NAB News Task," *Proc. ICASSP-96*.