Developments in Continuous Speech Dictation using the 1995 ARPA NAB News Task

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 benchmark test. In contrast to previous evaluations, the new Hub 3 test aims at improving basic SI, CSR performance on unlimited-vocabulary read speech recorded under more varied acoustical conditions (background environmental noise and unknown microphones). 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. 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. 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 order to encourage diversity in approaches, the test specifications provided no restrictions on the acoustic or language model training data used except that such data must predate August 1, 1995 (prior to the period from which the test data was taken), and the use of "stereo" sources of training data (with speech simultaneously recorded using both the standard close-talking microphone and other microphones) could be used only for training environmental compensation algorithms. 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 article boundaries and utterance order were known to the systems enabling the use of unsupervised transcription-mode adaptation.

The acoustic training data used by LIMSI includes a total of 46,146 sentences, 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.

For language modeling data, we used the newspaper texts and read speech transcriptions predating *July 30, 1995* (inclusive). This data includes the August'94 release of the CSR standard LM training texts distributed by LDC (years 88-94), the 1994 NAB development data (excluding the devtest data), the WSJ0/WSJ1 read speech transcriptions (85,343 sentences), and the 1994 and 1995 financial domain material (Hub 3 LM material). The texts from the last day *(31st of July, 1995)* were excluded in order to be able to extract from it a development test set for optimization of the LM and vocabulary list.

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. Acoustic modeling uses 48 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band every 10ms. Cepstral mean removal is performed for each sentence. The lexicon is represented using a set of 46 phones including silence. Each phone model is a tied-state left-to-right, CDHMM with Gaussian mixture ob-

servation densities (typically 32 components). The triphone contexts to be modeled are automatically 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. 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]. The word recognizer is then run using the set of models corresponding to the identified gender. Word recognition is performed in two steps, each with two passes.

- **Step 1:** 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 a gender-specific set 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:** 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].

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 with the log-normal approximation approach more parameters are typically used when more training data is available.

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.)

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

We used 65k bigram and trigram language models trained on 284M words of newspaper texts and the read speech transcriptions (85,343 sentences) predating *July 30, 1995* (inclusive). Texts containing about 17k words have been extracted from the *July 31st* texts to serve as development data (denoted dev95) according to the test text selection criteria determined by NIST (min and max lengths, manual verification for typos and readability). The LM training texts were cleaned to remove errors inherent to the texts or arising from processing with the distributed text processing tools. As done last year, the texts were transformed to be closer to the observed American 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. This year, we also processed the most frequent acronyms in the training texts in order to treat them as whole words instead of as sequences of independent letters. This processing resulted in 4% reduction of the test perplexity on the development texts.

The 65k word list was selected to minimize the OOV rate on the development texts, which resulted in selecting the most frequent words occuring in the WSJ texts from 92-94 (45M), the dev94 texts (1.9M) the WSJ0/1 transcriptions (1.4M), and the 1994 Hub3 texts (44M). 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. The lexical coverage on the dev95 test data is 99.4%. Perplexities and OOV rate are given in Table 1 for the dev95 texts and for the transcriptions of the ARPA Nov95 evaluation data (eval95). The trigram LM with backoff cutoffs of 0 and 1, contains 15.7M bigrams and 21.1M trigrams.

Test set	*dev95*	*eval95*
%oov	0.6	0.8
2-g px	222.2	239.3
3-g px	126.0	137.2

Table 1: OOV rate and perplexities for the dev95 and eval95 test texts.

Grammar condition	*Noise compens.*	*Speaker adapt.*	*% Word Error*	
			P0 data	*C0 data*
2-g	y	n	23.7	13.2
3-g	y	n	20.5	10.4
3-g	n	n	-	10.4
3-g	y	y	17.5	9.1
3-g	sw	y	17.5	8.6

Table 2: Word error rates on the ARPA Nov95 test data for different acoustic and language models: P0 and C0 denote respectively the Sennheiser data and the secondaries microphone data.

The 65k vocabulary contains 65,500 words and 72,637 phone transcriptions. A pronunciation graph is associated with each word so as to allow for alternate pronunciations. Frequent inflected forms have been verified to provide more systematic pronunciations.

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 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 are extracted from the North American Business (NAB) news texts during the 1-31 August 1995.

Table 2 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

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 3: 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.

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 2) is unchanged.[1] 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.

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 2).

In Table 3 the relative increase in word error for the P0 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.

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

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 are 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 LM training text materials so as to be able to model the most common acronyms as words, instead of as sequences of independent letters. The word error obtained on the multi-microphone P0 data was 17.5%. Environmental adaptation based on the $y = (x + n) * h$ model is demonstrated to be effective as it reduces 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 obtain 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 *et al.* "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.