Where Are We In Transcribing French Broadcast News?

J.L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, H. Schwenk

Spoken Language Processing Group (http://www.limsi.fr/tlp)
LIMSI-CNRS, B.P. 133, 91403 Orsay Cedex, FRANCE

ABSTRACT

Given the high flexional properties of the French language, transcribing French broadcast news (BN) is more challenging than English BN. This is in part due to the large number of homophones in the inflected forms. This paper describes advances in automatic processing of broadcast news speech in French based on recent improvements to the LIMSI English system. The main differences between the English and French BN systems are: a 200k vocabulary to overcome the lower lexical coverage in French (including contextual pronunciations to model liaisons), a case sensitive language model, and the use of a POS based language model to lower the impact of homophonic gender and number disagreement. The resulting system was evaluated in the first French TECHNOLANGUE-ESTER ASR benchmark test. This system achieved the lowest word error rate in this evaluation by a significant margin. We also report on a 1xRT version of this system.

1. INTRODUCTION

At LIMSI we started working on broadcast news (BN) transcription in 1996 [4] and have since developed models and systems for 7 languages: Arabic, English, French, German, Mandarin, Portuguese and Spanish [8]. Our experience is that with some reasonable amounts of data we can get comparable results in these language (under 20% for unrestricted broadcast news data). We have recently invested major effort in improving our BN transcription system for American-English [11], i.e. bringing the word error rate to around 10%. The main improvements come from better decoding, acoustic and language modeling, more effective adaptation, discriminative training, and better pronunciation models including pronunciation probabilities. In this work on BN transcription in French we found these general improvements to carry over in a relatively straight-forward manner. However, French speech recognition systems must address the relatively high lexical variety of the French language which results in large out-of-vocabulary (OOV) rates. Concerning lexical coverage, the number of distinct words in French must typically be double that of English in order to obtain the same word coverage under comparable conditions [6]. The inflectional morphology of German and its highly generative process of compounding lead to even lower lexical coverage for a given vocabulary size [9]. A large proportion of the observed lexical variety corresponds to homophones, which can be distinguished only the language model (LM) since there are no acoustic differences in how the words are pronounced. A comparative study of French and English showed that, given a perfect phonemic transcription, about 20% of the words in English newspaper texts are ambiguous, whereas 75% of the words in French newspaper texts have an ambiguous phonemic transcription [6]. This difference between French and English mainly stems from number and gender agreement for nouns, adjectives and past participles, and the high number of different verb forms which are often homophones [6]. Many of the observed automatic transcription errors come from the homophonic number and gender agreement, as in the language model training data many agreements are not observed. In order to address this problem we use a 200k word lexicon with multiple pronunciations to account for well-known phenomena such as liaison and mute-e; a larger text corpus to train the language models; and also investigated the use of a continuous space language model [13] as well as a part-of-speech (POS) based language model.

This updated system was evaluated in the TECHNOLANGUE-ESTER benchmark test for BN transcription systems in French [7].

2. LANGUAGE MODELS

For language modeling four different types of textual data were used. There are about 3.7M words of precise transcriptions of BN acoustic data (1998-2003), including almost 1M words of transcriptions of the acoustic training distributed for the ESTER evaluation. In addition, 89M words of rapid transcriptions of BN data dating from 1991-2001 were used. These texts were completed with a large amount of newspaper and newswire data (from 1987 to April 2004, 507M words), WEB newswire data (from 11/2003 to 04/2004, 23M words) and the newspaper texts (370M words) distributed for the ESTER evaluation. The fourth source is a list of about 200 journalist names from different French TV and radio sources

The test set of the ESTER evaluation consisted of 10 hours of radio broadcast news shows, taken from four stations occurring in the training data, one source (France Culture) without any transcribed training data and one unknown source. This test set was recorded from Octo-

	65k				200k				200k restricted			
	%OOV	2g	3g	4g	%OOV	2g	3g	4g	%OOV	2g	3g	4g
Dev ESTER	0.76	121.6	74.2	64.8	0.29	120.0	72.9	63.9	0.30	120.0	74.6	65.9
Dev LIMSI	0.95	151.8	91.1	79.2	0.40	149.2	89.1	77.8	0.58	163.4	101.4	89.6
Eval ESTER	1.01	138.1	89.7	80.4	0.31	135.0	87.2	78.5	0.38	141.6	94.1	85.2

Table 1: OOV rates and perplexity values for the 65k and the 200k word lists; the 2g, 3g, and 4g perplexities were calculated considering a theoretical word list of 1M words, in order to properly take into account the difference in the OOV rates.

ber to December 2004. Since the distributed development data contained only data from the four known sources and came from the period April to July 2003, we decided to design a more difficult development set to build less sharp but more robust language models. We added to the official development set (containing 97k words), data from various sources including French speaking African radio (19k words), France Culture (20k words), and excerpts from a variety of French TV and radio broadcasts (22k words). The resulting LIMSI development set includes a total of 158k words.

Two word lists were derived from the textual data described above: a classical 65k list and a larger 200k list in order to increase the lexical coverage. Both word lists were optimized using the criterion described in [2] to select the vocabulary which minimizes the OOV rate on the LIMSI development set.

Text normalization has a impact on both the OOV rate and on the LM perplexity. Previous work [1] on the French language showed that case sensitivity, may marginally reduce the lexical coverage, but leads to more performant LMs. Thus, despite the fact that the ESTER scoring was case insensitive, we built case sensitive language models.

The training data was split into 5 text subsets and 4-gram back-off LMs were built for these sources using the SRI Toolkit [12] with the modified Kneser-Ney discounting method and then interpolated to get the final LM. The text subsets and interpolation coefficients, the latter determined by the EM algorithm, were chosen so as to minimize the perplexity on the LIMSI development set. Another set of n-gram back-off LMs were built using only the text data distributed in the ESTER evaluation, and a 200k word list was selected with the same restriction (200k restricted), in order to measure the impact of the additional data available at LIMSI. The OOV rates and the perplexity values on the different development and evaluation sets are given in Table 1.

Although the detailed transcriptions of the audio data represent only a small fraction of the available data, they have an interpolation coefficient of 0.43. This shows clearly that these detailed audio transcriptions are the most appropriate text source for the task. Given this, a continuous space neural network LM [13] was trained on the transcriptions, the neural network being used to simultaneously learn the projection of the words in a continuous space and to estimate the n-gram probabilities. The neural network LM is interpolated with the baseline back-off LM.

To reduce ASR errors due to homophonic number and gender disagreement, a specific rescoring scheme is applied in the last decoding step. Given an ASR hypothesis, a lattice of homophones is generated and rescored using a morphosyntactic class-based LM interpolated with the baseline word n-gram LM. The rescoring process is performed by weighting the edges only with the linguistic score and a consensus decoding step yields to the final hypothesis.

3. PRONUNCIATIONS

Generating pronunciations for the 200k word list was a challenge since this work usually requires some manual intervention. There are two major sets of words that were added when increasing the vocabulary from 65k to 200k words. The first set corresponds to various forms of verbs and does not pose any particular problem from the pronunciation point of view, other than it being important to include alternate pronunciations for reduced forms. The second set consists of mainly proper names, many of which are foreign names for which pronunciations are not properly predicted by a grapheme-to-phoneme tool. Since the pronunciation of such foreign proper names depends on the speaker's knowledge of the original language, we have tried to include multiple forms to represent both how a native French and someone familiar with the language would speak the name.

The basic pronunciations are taken from the LIMSI French lexicon, and make use of a 35-phone set (3 of which are used for silence, filler words, and breath noises). Baseform pronunciations for the missing words are generated using a grapheme-to-phoneme conversion tool, and alternative pronunciations are added semi-automatically. The most frequent inflected forms have been verified to provide more systematic pronunciations. In particular, liaison consonants, which are handled by using contextual variants in the pronunciation dictionary, were added to determinants and to the frequent common nouns and adjectives [3]. These variants are important to represent explicitly in the lexicon, since a fixed HMM structure is used. For French this is especially important to handle schwa (mute-e) insertion and deletion, which can arise from different regional speaking styles as well as as from hesitations. Another problematic situation arises when phonemes are missing, as often occurs in casual speech.

Such sequential reductions, similar to contracted forms in English are frequent in French. Other shortened pronunciations are not reflected in writing: *il y a* ("there is") is often uttered as *y a*, and the word *cette* ("this") may be realized as /sEt/, /sEtx/(C.) or /st/(V).

The pronunciation probabilities are estimated from the observed frequencies in the training data resulting from forced alignment, with a smoothing to account for unobserved pronunciations. The 200k word lexicon has 276k pronunciations.

4. SEGMENTATION

The LIMSI segmentation and clustering is based on an audio stream mixture model [5]. 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.

This procedure and the associated models developed for American English BN data has been used as is for all the language we have worked on without any need for adaptation. We however found that the initial segmentation based on GMM modeling speech, noise and music can be slightly improved by adapting the models to the targeted data. The main variability is coming from the music segments and the speech over music which are clearly source dependent. We therefore adapted this modeling using some music and segment segments from the various sources. The overall gain in WER due to this adaptation is in fact quite small (about 0.1%).

5. ACOUSTIC MODELS

The speech features consist of 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band (or 0-3.8kHz 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. LPC-based cepstrum coefficients are then computed. These cepstral coefficients are normalized on a segment cluster basis using cepstral mean removal and variance normalization. Each resulting 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. This feature vector is linearly transformed (MLLT) to better fit the diagonal covariance Gaussians used for acoustic modeling.

The acoustic models were trained on about 190 hours of BN training data. Instead of using the manual segmentations in the reference transcriptions, the words were aligned with the automatically determined segments created by the audio partitioner. This significantly simplifies our training procedure and is coherent with the subsequent decoding.

For the final decoding pass, the acoustic models include 23k position-dependent triphones with 12k tied states, obtained using a divisive decision tree based clustering algorithm with the 35 phones. Two sets of MLLT-SAT gender-dependent acoustic models were built for each data type (wideband and telephone) using MAP adaptation of SI seed models and MMI training.

6. DECODING

Decoding is performed in three passes, where each pass generates a word lattice which is expanded with a 4-gram LM. Then the posterior probabilities of the lattice edges are estimated using the forward-backward algorithm and the 4-gram lattice is converted to a confusion network with posterior probabilities by iteratively merging lattice vertices and splitting lattice edges until a linear graph is obtained. This last step gives comparable results to the edge clustering algorithm proposed in [10]. The words with the highest posterior in each confusion set are hypothesized.

For the first and second passes the lattice rescoring step is done using a standard 4-gram language model, while in the third pass rescoring is done with the neural network model and the POS language model. Also to speedup the third pass the search space for each audio segment to be decoded is restricted to the a word graph derived from the lattice generated in the second pass. This results in a decoding speed of about 7.5xRT (0.1xRT for the segmentation, 1.2xRT for the first pass, 4.6xRT for the second pass, and 1.6xRT for the third pass). Unsupervised acoustic model adaptation is carried out for each speaker between the decoding passes making use of the hypotheses of the previous pass. This done by means of a constrained MLLR adaptation followed by a unconstrained MLLR. For the regular MLLR adaptation, two regression classes (speech and non-speech) are used in the second pass, whereas a data driven clustering with a variable number of classes is used in the third pass.

A real-time version of the decoding procedure has also been implemented. For this condition the decoding is reduced to two passes with very tight pruning thresholds (especially for the first pass) and with fast Gaussian computation based on Gaussian short lists. The 1xRT time budget is divided as follows: 0.08xRT for segmentation, 0.14xRT for the first decoding step, and 0.78xRT for the

Audio	Texts	Vocab.	LM	RT	WER
all (90+ 100)	all	65k	4g		11.0
	all	200k	4g		10.7
	all	200k	4g, NN		10.4
	all	200k	4g, POS		10.5
	all	200k	4g, NN, POS	7.5	10.3
90h	all	200k	4g, NN, POS	7.5	10.8
90h	restrict	200k^r	4g, NN, POS	7.5	11.6
all	all	65k	4g, NN, POS	1.1	13.7

Table 2: Word error rates for different training configurations: acoustic training data (all vs 90h ESTER data), word lists (65k, 200k and 200k restricted), LM training data (all vs. ESTER texts), and language modeling (4g word LM (4g), with neural LM (4g,NN), and with class LM rescoring (4g,NN,POS))

second decoding step. The compute platform is an Intel Pentium 4 extreme (3.2GHz, 4GB RAM) running Fedora Core 2 with hyperthreading.

7. EXPERIMENTAL RESULTS

The experimental results on the ESTER development data are given in Table 2. It can be seen by comparing the first two rows of the table that the gain obtained with the 200k language model relative to the 65k LM is quite limited. The absolute error rate reduction is 0.3% which is less than the OOV rate reduction of about 0.5% when going from 65k to 200k. A larger reduction could have been expected given that an OOV word usually generates 1.5 to 2 errors. Adding the neural network language model reduces the word error rate from 10.7% to 10.4%, whereas the POS language model reduces the word error rate by about 0.2% from the same baseline. Using both the NN and POS LMs reduces the word error to 10.3%. Given the small gain obtained with the 200k language model, the 65k LM was used for the 1xRT system which has a WER of 13.7% on the same development data (last line in Table 2).

The WER increases to 10.8% when using only 90h of acoustic training data, i.e. by removing the 100h of data from 1994 to 1999. The loss is limited certainly because this data is not very representative of the sources used in the ESTER evaluation. Using only the restricted text condition to train the language model further increases the word error to 11.6%.

On the 10 hours of the ESTER evaluation data, the full 200k system obtained a word error rate of 11.9%, achieving the lowest word error rate in this evaluation by a significant margin.

8. CONCLUSIONS

This paper has described recent improvements in processing broadcast news speech in French. Advances made in our English broadcast news system [11] have been incorporated in the system and were found to give comparable improvements. To address the relatively high lexical variety and homophone rate in the French language, a 200k word back-off n-gram case-sensitive language model is used, as well as a continuous space language model and a part-of-speech language model are used in a final rescoring pass. A significant effort was devoted to extending the pronunciation lexicon from 65k to 200k words and to add alternative contextual pronunciations to more systematically model liaisons and mute-e. The resulting word error rate on the ESTER development data is 10.3%, which is in the same range as results obtained on English despite the high flexional properties of the French language and the large number of homophones. This system was also evaluated in the first French ESTER ASR benchmark test.

REFERENCES

[1] Adda, G., Adda-Decker, M., Gauvain, J.L., and Lamel, L., ”Text normalization and speech recognition in French”, *EuroSpeech'97*, **5**:2711-2714, Rhodes, Sept 1997.

[2] Allauzen, A., and Gauvain, J.-L., ”Construction automatique du vocabulaire d'un système de transcription”, *Journées d'Etude sur la Parole 2004*, Fes, 2004.

[3] Boula de Mareuil, P., Gendner, V., Adda-Decker, M., “Liaisons in French: a corpus-based study using morpho-syntactic information,” *ICPhS*, Barcelona, Aug 2003.

[4] Gauvain, J.L., Adda, G., Lamel, L., and Adda-Decker, M., “Transcribing broadcast news: The LIMSI Nov96 Hub4 system,” *ARPA Spoken Language Technology Workshop*, 56-63, Chantilly, Virginia, Feb 1997.

[5] Gauvain, J.L., Lamel, L., and Adda, G., “Partitioning and Transcription of Broadcast News Data,” *ICSLP'98*, 1335-1338, Sydney, Dec. 1998.

[6] Gauvain, J.L., Lamel, L.F., Adda, G., Adda-Decker, M., “Speaker-Independent Continuous Speech Dictation,” Speech Communication, **15**(*), Sept. 1994.

[7] Gravier, G., Bonastre, J.F., Galliano, S., Geoffrois, E., Mc Tait, K., and Choukri, K., “The ESTER evaluation campaign of Rich Transcription of French Broadcast News,” *LREC'04*, Lisbon, May 2004.

[8] Lamel, L., and Gauvain, J.L.,“Automatic Processing of Broadcast Audio in Multiple Languages,” *Eusipco02*, Toulouse, Sep 2002.

[9] McTait, K., and Adda-Decker, M., “The 300k LIMSI German Broadcast News Transcription System,” Processing of Broadcast Audio in Multiple Languages,” *Eurospeech'03*, Geneva, Sept 2003.

[10] Mangu, L., Brill, E., and Stolke, A., “Finding Consensus Among Words: Lattice-Based Word Error Minimization,” *Eurospeeech'99*, 495-498.

[11] Nguyen, L., et al., “The 2004 BBN/LIMSI 10xRT English Broadcast News Transcription System,” *DARPA RT'04 workshop*, Palisades, Nov. 2004.

[12] Stolcke, A., ”SRILM - An Extensible Language Modeling Toolkit”, *ICSLP'02*, Denver, Sept 2002.

[13] Schwenk, H., and Gauvain, J.L., “Neural Network Language Models for Conversational Speech Recognition,” *ICSLP'04*, Jeju, Oct. 2004.