Unsupervised model adaptation on targeted speech segments for LVCSR system combination

Richard Dufour, Fethi Bougares, Yannick Estève, Paul Deléglise

LIUM - University of Le Mans
Avenue Laennec, 72085 Le Mans, France
firstname.lastname@lium.univ-lemans.fr

10.21437/Interspeech.2010-297

Abstract

In the context of Large-Vocabulary Continuous Speech Recognition, systems can reach a high level of performance when dealing with prepared speech, while their performance drops on spontaneous speech. This decrease is due to the fact that these two kinds of speech are marked by strong acoustic and linguistic differences. Previous research works have been done to detect and repair some peculiarities of spontaneous speech, such as disfluencies, and to create specific models to improve recognition accuracy: a large amount of data—which is expensive to collect—is needed to see improvements. In this paper, we present a solution to create specialized acoustic and language models, by automatically extracting a data subset from the initial training corpus containing spontaneous speech, and adapting initial acoustic and linguistic models on it. As we assume these models can be complementary, we propose to combine general and adapted ASR system outputs. Experimental results show statistically significant gain, at a negligible cost (no additional training data and no human intervention).

Index Terms: spontaneous speech detection, model adaptation, ASR system combination.

1. Introduction

In Broadcast News (BN), different kinds of speech can be found, including prepared speech (close to read text) and spontaneous speech (interviews, debates, dialogues...). Some Large-Vocabulary Continuous Speech Recognition (LVCSR) systems are designed to transcribe speech from large audio sources, by using specialized models to deal with all kinds of speech that can emerge. As pointed in many papers, for example [1, 2], differences exist between kinds of speech, such as, for the spontaneous one, ungrammatically, different language register or disfluences (filled pause, repetition, false start...), and can appear at an acoustic and linguistic level. Various studies have focused on the detection and the correction of spontaneous speech evidence, such as disfluencies [3, 4] or ungrammaticality [5]. These studies led to new research works on detection of spontaneous speech [6], to automatically find spontaneous speech segments and feed them to an ASR system.

Higher Word Error Rate (WER) values are obtained by state-of-the-art Automatic Speech Recognition (ASR) systems when transcribing data containing spontaneous speech [7]. Specifically modeling several components of an ASR system to process spontaneous speech could be a solution to improve WER. For example, in [8], the authors focused on spontaneous speech for pronunciation modeling, and found a possible ASR improvement on the spontaneous speech portion of their BN corpus by taking into account specific phonetic observations. Moreover, Furui in [7] presents advances in spontaneous speech recognition, and shows that word accuracy, when processing spontaneous speech, is better when acoustic and language models are trained with a spontaneous speech corpus instead of a read-speech corpus. These various results confirm that creating a general model to cope with all kinds of speech is unrealistic, since differences between them are too large.

In this paper, we are interested in dealing with spontaneous speech occurring in BN. Specifically, we seek to improve LVCSR performance by using specific models and combining recognition hypotheses from specialized ASR systems: with each system using a different knowledge base, we expect that they can propose complementary hypotheses. This potential complementarity should permit to improve LVCSR performance [9].

Getting a large amount of data is necessary to create reliable models [10], but collecting data is difficult and expensive. Generally, the amount of training data available is not sufficient to create such specialized models. Adapting general models with some specific data is the solution adopted by [11]. The limit of this method is that additional data is needed to adapt acoustic and linguistic models, which have to be specialized for a targeted domain (baseball radio speech).

Our proposed approach follows the idea of adapting models used in ASR on a data subset extracted from the initial training corpus of the general acoustic and linguistic models without data addition or human involvement.

Because no data will be provided, the main difficulty is to automatically extract data from our training corpus to adapt models to spontaneous speech. We use a tool developed during previous spontaneous speech studies [6] in order to automatically find spontaneous speech segments. This automatic spontaneous speech detection tool is able to extract spontaneous speech segments on which acoustic and linguistic model adaptation will be made. Finally, a combination of recognition hypotheses from the general ASR system and the spontaneous-targeted ASR system will be made.

2. Overview of the LIUM ASR system

Based on the CMU Sphinx decoders, the LIUM ASR system [12] is a multi-pass system developed to process French Broadcast News audio recordings. It ranked second during the ESTER 2005 French evaluation campaign [13], and was the best open source system participating in the ESTER 2 evaluation campaign [14] in 2008. This section briefly describes the

This research was supported by the ANR (Agence Nationale de la Recherche) under contract number ANR-09-BLAN-0161

LIUM ASR system components.

2.1. Decoding

The decoding process involves 5 passes, using the Sphinx3.7 decoder during the first and second passes, and the Sphinx4 decoder from pass 3 to pass 5:

1. The first pass is performed with a trigram language model and acoustic models corresponding to the gender and the bandwidth detected by the segmentation process. Acoustic models and recognition hypotheses allow to compute a CMLLR (Constrained Maximum-Likelihood Linear Regression) transformation for each speaker.
2. The second pass applies a CMLLR transformation with the best hypothesis generated during the first pass. Although the same language model is used, acoustic models are replaced by models trained with SAT-CMLLR (Speaker Adaptive Training) adaptation. This pass generates word-graphs.
3. In the third pass, the word-graphs are used to drive graph-decoding with full 3-phone context, which enables a better acoustic precision, particularly in inter-word areas. This pass generates new word-graphs.
4. The fourth pass consists in rescoring the word-graph generated during the third pass with a quadrigram language model.
5. The last pass generates a confusion network from the last word-graphs, and applies the consensus method to extract the final one-best hypothesis. This pass provides word posteriors as confidence measure values.

2.2. Acoustic models

During training of the general model, several models are created. Bandwidth dependent-models (wideband / narrowband) composed of 6500 tied states are created first, and are then mapped to generate gender-dependent models; as a result, four specialized models are obtained: Male-Wideband (M_WB), Female-Wideband (F_WB), Male-Narrowband (M_NB) and Female-Narrowband (F_NB). These models are used to compute CMLLR transformation matrices to obtain SAT-CMLLR models with 7500 tied states.

2.3. Language Model

Data used to build the language model come from manual transcriptions of BN (used to train acoustic model), newspaper articles and web resources.
To build the vocabulary, we generate a unigram model as a linear interpolation of unigram models trained on the various training data sources listed above. Then, we extract the 122k most probable words from this language model. Using this vocabulary, all the textual data of the training corpus is used to train trigram and quadrigram language models. To estimate and interpolate these models, SRILM is employed using the modified Kneser-Ney discounting. No cut-off is applied on unigrams, bigrams, trigrams and quadrigrams. The models are composed of 121k unigrams, 28M bigrams, 160M trigrams, and 371M quadrigrams.

3. Automatic adaptation of models

In this section, we describe how spontaneous speech is automatically detected, and how acoustic and language models are adapted to this kind of speech.

3.1. Automatic spontaneous speech detection

The automatic detection follows the method developed in [6]. This statistical method uses a classifier dealing with prosodic and linguistic features computed on a corpus composed of 11 files containing French BN data, for a total duration of 11h37' and a total number of 11,821 segments. This corpus was manually labeled by human judges, according to three classes of spontaneity (*prepared*, *low spontaneous* and *high spontaneous speech*). A nearly equivalent number of segments has been manually labeled for each class.
In order to perform the detection process, audio files must be cut into segments using the LIUM automatic segmentation and diarization system [15]. Then for each segment, we compute acoustic features using LIUM ASR system and linguistic features from the reference transcriptions. All extracted features for each individual segment are given to a classifier, to determine a class of spontaneity. The classification tool used is *icsiboost*, an open source tool based on the *AdaBoost algorithm* like the *Boostexter* software [16]. It is a large-margin classifier based on a boosting method of weak classifiers. Finally, weighted finite-state transducers are used to re-estimate classification probabilities obtained on each segment, by taking into account classification results on the previous and the next segment.
Manually labeled segments will constitute our training dataset for classification. Then the automatic detection of spontaneous speech segments will determine which segments to extract on the training corpus. The detection tool reaches a precision of 69.3% and a recall of 74.6% on spontaneous speech segments.

3.2. Model adaptation

The most direct way to produce a spontaneous speech recognition system is to estimate appropriate models by type of speech [10]. However it is not easy to get enough data to create specialized models. Thus, instead of creating specialized models, we propose to adapt our models with the subset of spontaneous speech automatically extracted from the initial training corpus.

3.2.1. Acoustic model adaptation

Figure 1 shows the acoustic model adaptation process. As the extracted corpus of spontaneous speech is not sufficient to create global acoustic models, the two first decoding passes use general acoustic models already used in the ASR system. So as general models are already adapted by gender and bandwidth (pass 2), a MAP adaptation is then applied on each gender/bandwidth model with the spontaneous speech subset data (divided by gender/bandwidth condition). As a result, we obtained four spontaneous speech specific models depending on gender and bandwidth. Decoding from pass 3 takes three times real-time, and ten times real-time with all the passes.

3.2.2. Language model adaptation

The initial trigram and quadrigram language models were estimated using seven different training textual corpus: five came from different newspapers, one from web data, the last one was composed by the manually transcribed annotations associated to the audio files used to train the acoustic models. The spontaneous speech detection tool extracted spontaneous speech segments from the last training corpus presented above: manual

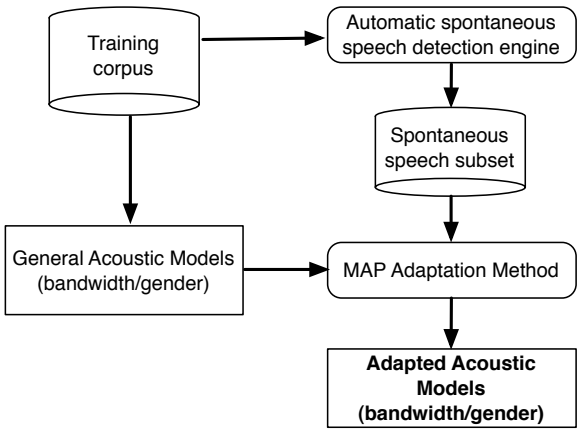

Figure 1: Adaption of acoustic models on spontaneous speech.

transcriptions associated to these spontaneous segments were used as an eighth training corpus. These eight training corpora were used to estimate trigram language models and quadrigram language models. The trigram language models were linearly combined in order to produce the new trigram language model used in passes 1 to 3 into the LIUM ASR system. The quadrigram LM used in pass 4 was estimated following the same way. The linear coefficients were optimized on the manual transcriptions associated to the spontaneous speech segments detected in a development corpus.
So the main difference between initial language models and spontaneous speech specialized language models is the development corpus used to optimize linear weights (the entire development corpus for the initial models, only the spontaneous speech segments for the specialized models). Moreover, a reinforcement of the observations occurring on spontaneous speech segments in the training corpus was made by using the manual transcriptions associated to these segments as a specific corpus. Indeed, these segments were already present in the initial training corpus used to train acoustic models.

4. Corpus

This section describes the corpus used to train general models, including the amount of data used for spontaneous speech adaptation (subset of the general train corpus).
To build acoustic models, large speech corpora are essential. The training corpus used is composed of 240 hours provided by the ESTER 1 & 2 evaluation campaigns (mainly prepared speech, described in [12]), plus 80 hours of transcribed French radiophonic shows provided by the EPAC project[17] (mainly spontaneous speech).
To build specialized acoustic models, a part of the training corpus had been extracted thanks to automatic spontaneous speech detection. This subset, which represents 133 hours of spontaneous speech, will be used to adapt the general acoustic models. As acoustic models are adapted depending on gender/band, table 1 presents the duration of training data (in hours) used to train the global acoustic model and the adapted acoustic model to spontaneous speech, according to gender and bandwidth.

Table 1: Duration (in hours) of acoustic training for the global and the spontaneous adapted models (extracted) according to gender and bandwidth.

Corpus	Global	Extracted
M_WB	161	84
F_WB	53	21
M_NB	33	23
F_NB	9	5

5. Experiments

The experiments are carried out using the official development and test corpus of the ESTER 2 campaign and the EPAC project. The development corpus contains 13 hours, and the test corpus contains 16 hours of audio recording.

5.1. System analysis

In order to examine the impact of the adapted models, we carried out speech decoding experiments. The basic idea is to do two decodings with the LIUM ASR system, using the basic models and the adapted ones separately. Firstly, we want to see the impact of both models on highly spontaneous speech segments. Table 2 presents WER over the development corpus on these segments only, using the baseline and the adapted system. Results are shown by bandwidth and gender.

Table 2: WER on high spontaneous segments on development corpus using baseline and adapted system.

High Sponta.	Baseline	Adapted
M_WB	28.0	27.3
F_WB	25.0	26.2
M_NB	31.2	30.6
F_NB	27.5	28.4
Global	27.1	27.0

We can see that the global WER changes on spontaneous speech segments, with a very slight decrease when using the adapted system. It is also interesting to note that results are different depending on gender and bandwidth: gains are obtained on male speakers, while losses are obtained on female speakers. The relative low gain on spontaneous speech segments could be explained by the difference between features used into automatic spontaneous speech detection system on training corpus, and on development corpus, with respectively the use of reference transcriptions and ASR transcriptions. Moreover, disparate results between gender and bandwidth could be due to different amount of adaptation data, as we can see in table 1.
For all these reasons, we focus our efforts on combining baseline and adapted systems.

5.2. System combination

We have shown that using only an adapted system is not the best way to decrease WER. But according to hypotheses, it seems that a better word hypothesis could be introduced by the adapted system. In this context, a combination of ASR outputs might be a solution to get the best word hypothesis from each system.

Since the LIUM ASR system can provide confidence measures (CM), estimated from acoustic and linguistic scores, and associated to each decoded word, CM could be a good indicator to choose hypothesis. For these reasons, we combined baseline and adapted system outputs using the ROVER method [18]. This method seeks to reduce WER by hybridating different ASR outputs, where confidence measures are taken into account.
Firstly, we minimized the WER of the composite ASR outputs on development corpus. As we want to use confidence measures of combined ASR outputs, the @ parameter (frequency of word occurrences) will not be set to 1, but can take any value between 0 and 0.9 since only two systems are combined. Smaller WER have been obtained by setting *Conf(@)* to 0.8. Table 3 presents results obtained with ROVER combination, and the minimum WER that we could expect if the best word hypothesis is always chosen (oracle).

Table 3: ASR performance (WER and relative gain %) comparing baseline, ROVER combination and oracle computing.

	Development	Test
Baseline	21.08	18.85
Adapted⊕Baseline	20.67 (-1.94%)	18.53 (-1.70%)
Oracle	18.52 (-12.1%)	16.47 (-12.6%)

Combination of outputs show the complementarity of our systems, as this combination can reach an accuracy improvement of 1.7% on our test corpus. Even if the gain is minimal for now, the oracle shows high potential gain (12.6%) with an ideal combination. Then, inter-system comparisons (baseline and adapted) was computed with the NIST *sc_stat* tool, by doing a Matched Pairs Sentence-Segment Word Error (MAPSSWE) test. It indicates that the improvement with system combination is statistically significant at the level of **p=0.001**.

6. Conclusion

Assuming that spontaneous speech significantly degrades performance of ASR systems, many previous works have concentrated substantial efforts on collecting spontaneous speech data (with human transcribers) in order to create specialized acoustic and linguistic models. Because getting a sufficient amount of data to create such models is difficult, other strategies must be considered to deal with spontaneous speech.
In this article, we presented a method that seeks to adapt acoustic and linguistic models on spontaneous speech. The main idea is not to add specific data, but to use slightly differently training data already used in ASR systems. The proposed adaptation method is fully automated, the subset corpus of spontaneous speech is extracted with an automatic spontaneous speech detection. This subset is then used to adapt acoustic models , and to interpolate the general language model with extracted transcriptions.
Experiments shows that the use of different models, estimated on the same training data but adapted differently, has a positive impact on the word error rate when they are used complementarily by using the ROVER method. A small but statistically significant reduction of the word error rate has been observed, with a relative gain of 1.7% on the test corpus. Moreover, this method does not need human expertise or human transcribers to get specific data, while having a relatively low extra decoding cost, specially if parallel decoding is performed.
As future work, we will have to focus our efforts in word hypothesis choices, by using more sophisticated methods, such as Confusion Network Combination. Indeed, the oracle gain is of 12.6% so a lower word error rate should be reached. Otherwise, spontaneity scores given by detection system can also be used in the final hypothesis choices.

7. References

[1] M. Nakamura, K. Iwano, and S. Furui, "Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance," vol. 22, no. 2, 2008, pp. 171–184.

[2] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," in *Interspeech*, Lisbon, Portugal, 2005, pp. 1781–1784.

[3] Y. Liu, E. Shriberg, A. Stolcke, and M. Harper, "Comparing HMM, Maximum Entropy, and Conditional Random Fields for Disfluency Detection," *Interspeech*, pp. 3313–3316, 2005.

[4] M. Lease, J. M., and E. Charniak, "Recognizing Disfluencies in Conversational Speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1566–1573, 2006.

[5] P. de Mareüil, B. Habert, F. Bénard, M. Adda-Decker, C. Barras, G. Adda, and P. Paroubek, "A quantitative study of disfluencies in French broadcast interviews," *Workshop Disfluency In Spontaneous Speech (DISS)*, 2005.

[6] R. Dufour, Y. Estève, P. Deléglise, and F. Béchet, "Local and global models for spontaneous speech segment detection and characterization," in *ASRU*, Merano, Italy, 2009.

[7] S. Furui, "Recent advances in spontaneous speech recognition and understanding," in *IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 1–6.

[8] E. Fosler-Lussier, S. Greenberg, and N. Morgan, "Incorporating contextual phonetics into automatic speech recognition," in *International Congress of Phonetic Sciences*, San Francisco, USA, 1999, pp. 611–614.

[9] D. Ellis, "Improved recognition by combining different features and different systems," in *The 19th Annual AVIOS Conference*, San Jose, CA, USA, 2000.

[10] S. Furui, M. Nakamura, T. Ichiba, and K. Iwano, "Why is the recognition of spontaneous speech so hard?" in *Lecture notes in Computer Science*, 2005, pp. 9–22.

[11] Y. Ariki, T. Shigemori, T. Kaneko, J. Ogata, and M. Fujimoto, "Live Speech Recognition in Sports Games by Adaptation of Acoustic Model and Language Model," in *Interspeech*, Geneva, Switzerland, 2003, pp. 1453–1456.

[12] P. Deléglise, Y. Estève, and T. Merlin, "Improvements to the LIUM French ASR system based on CMU Sphinx: what helps to significantly reduce the word error rate?" in *Interspeech*, Brighton, UK, 2009.

[13] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of French broadcast news," in *Interspeech*, Lisbon, Portugal, 2005.

[14] S. Galliano, G. Gravier, and L. Chaubard, "The ester 2 evaluation campaign for the rich transcription of french radio broadcasts," in *Interspeech*, Brighton, Royaume-Uni, Septembre 2009.

[15] S. Meignier and T. Merlin, "LIUM_SpkDiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, Dallas, Texas, USA, 2010.

[16] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," vol. 39, 2000, pp. 135–168.

[17] Y. Estève, T. Bazillon, J.-Y. Antoine, F. Béchet, and J. Farinas, "The EPAC corpus: manual and automatic annotations of conversational speech in French broadcast news," in *LREC*, Malta, 2010.

[18] J. G. Fiscus, "A Post-Processing System To Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," in *ASRU*, Santa Barbara, USA, 1997, pp. 347–352.