INCORPORATING MLP FEATURES IN THE UNSUPERVISED TRAINING PROCESS

Thiago Fraga-Silva [1], Viet-Bac Le [2] , Lori Lamel [1], Jean-Luc Gauvain [1]

[1] Spoken Language Processing Group
LIMSI - CNRS B.P. 133, 91403 Orsay FRANCE
{thfraga,lamel,gauvain}@limsi.fr

[2] Vocapia Research
3, Rue Jean Rostand, 91400 Orsay FRANCE
levb@vocapia.com

ABSTRACT

The combined use of multi layer perceptron (MLP) and perceptual linear prediction (PLP) features has been reported to improve the performance of automatic speech recognition systems for many different languages and domains. However, MLP features have not yet been used on unsupervised acoustic model training. This approach is introduced in this paper with encouraging results. In addition, unsupervised language model training was also investigated for a Portuguese broadcast speech recognition task, leading to a slight improvement of performance. The joint use of the unsupervised techniques presented here leads to an absolute WER reduction up to 3.2% over a baseline unsupervised system.

***Index Terms*—** Unsupervised Training, MLP features, Acoustic Modeling, Language Modeling

1. INTRODUCTION

Acoustic (AM) and Language Models (LM) are two main components of any Large Vocabulary Continuous Speech Recognition (LVCSR) system. Usually, training these models requires large amounts of data to achieve suitable performance levels. In practice, for acoustic modeling, it implies the need to transcribe hundreds of hours of speech. However, obtaining manual transcriptions is an expensive and time-consuming task. For languages that have a small number of speakers, it might be even harder, since engaging and training "annotators" becomes an arduous step in the transcription process. On the other hand, language models are usually estimated using millions of words. For a number of languages and domains, text data can be gathered from the Web. However, it is known that audio transcriptions play a major role in language modeling, helping to generate better LM estimates.

To alleviate the need of manual transcriptions during system development, unsupervised training methods can be used. Unsupervised acoustic model training (AM-UT) is a technique that has been gaining popularity over the last years and has been successfully applied to different languages [1, 2, 3, 4, 5] and domains, such as Broadcast News [2, 6, 7], Broadcast Conversations [2] and Conversational Telephone Speech (CTS) [8, 9]. Notwithstanding, few results have been reported concerning unsupervised language model training (LM-UT) [9] or adaptation [10].

Most of the published experiments on AM-UT make use of Hidden Markov Models (HMMs) based on raw features, such as perceptual linear prediction (PLP) [11] features, extracted directly from the audio stream. On the other hand, the combined use of PLP and multi layer perceptron (MLP) features have consistently been reported to improve LVCSR performances when the models are trained in a supervised manner [12, 13, 14]. Despite that, no studies have been reported in which unsupervised training methods have been applied to estimate the MLP parameters. In this paper, we propose to address this issue. In comparison to PLP models, the unsupervised training of MLP-based models introduces a new source of uncertainty. In addition to the HMMs, it is also necessary to estimate the parameters of the MLP neural networks that are used during feature extraction. The topic of LM-UT is also addressed in this paper. It is shown that the use of automatic transcriptions lead to a small gain of performance in language modeling on a broadcast recognition task. Since manual transcriptions were available, the unsupervised systems were compared with their equivalent supervised ones. Similar comparisons were performed on a CTS recognition task using PLP features [8, 9]. Such analysis helps to measure how much of performance is lost when labelled data is not available, and to identify possible weaknesses of the unsupervised training approach.

This work was performed using a system developed for Portuguese, which is one of the languages with the highest number of speakers in the World, but has received less attention in the community than other languages like English, Arabic or Mandarin. The unsupervised approaches assessed here can be applied to any other language, since the focus is on the impact of audio transcriptions on system development.

This paper is presented as follows. Section 2 describes the task and system in which the experiments were carried out. The following section describes the experiments realized in LM-UT. Section 4 is dedicated to AM-UT with PLP and MLP features. A final discussion and conclusions are presented in Section 5.

This work has been partially supported by OSEO, the French State agency for innovation, under the Quaero program.

2. TASK AND SYSTEM OVERVIEW

The experiments were carried out using the LIMSI speech recognition toolkit, with acoustic, lexical and language models developed for Portuguese and tested with broadcast data.

2.1. Corpora

The manually transcribed training data used in these experiments were collected under the Quaero Programme [1]. This corpus, henceforth referenced as *trn11*, contains about 56 hours of speech of shows broadcast between June and December 2010. The amount of data is roughly equally shared between Broadcast News and Broadcast Conversations. The development (*dev*) set contains about 3.5 hours of data with manual reference transcriptions and consists of shows broadcast on January 2011.

The language model training data include about 640 million words from nine different written sources, such as newspapers, newswires and blogs. These data cover the period from 1991 to 2010. About 30k words of transcriptions from RTP shows broadcast in 2000 were also used. The manual transcriptions of the *trn11* set used on the estimation of the supervised models contain about 560k words. The text data was normalized in order to convert numerical forms (cardinal, ordinal, date, currency...) and abbreviations to spoken forms. A 3-gram casing LM, trained from pre-selected text sources, was applied to all texts, except the *trn11* transcriptions, for which it was assumed that the correct case was already assigned.

2.2. System description

The system used in these experiments is quite similar to the system described in [15]. It makes use of n-gram language models and acoustic models based on continuous density HMMs. Each phone is modeled by a tied-state left-to-right context-dependent triphone HMM, with Gaussian mixture observation densities. PLP, pitch and MLP features were used. The PLP feature vector contains 39 components, including 12 cepstrum coefficients and log energy with their first and second derivatives. In the supervised PLP-based system, a 3-dimensional pitch feature vector (pitch with first and second derivatives) is added to the original PLP feature resulting in a vector with 42 components (PLP+F0). The pitch features were also used in all the MLP-based systems. The MLP feature vector has 39 components extracted from the hidden layer of a bottleneck MLP network (see Section 4.2 for details). The phone set contains 35 phones, as well as special units for silence, breath and hesitation markers.

The vocabulary was automatically selected based on interpolation of unigram LMs: 1) one unigram model was trained for each text source; 2) these LMs were interpolated so as to minimize the perplexity on the *dev* set; 3) a 65k-word vocabulary was selected from the highest probability unigrams. The out-of-vocabulary rate observed on the *dev* set was 1.1%.

Component language models were estimated from each of the different sources using the vocabulary selected and interpolated modified Kneser-Ney smoothing. Therefore, 2-, 3- and 4-gram LMs were built by interpolation of these component models with weights automatically chosen in order to minimize the perplexity on the *dev* set.

A pronunciation dictionary was obtained for the vocabulary via a ruled-based grapheme to phoneme (G2P) converter. This module has a pre-processing step which performs syllabification and stress syllable marking before doing the G2P conversion. About 530 rules are used to generate (with few exceptions) a unique pronunciation for each word. Some pronunciation variants were manually added for a few frequent words. Alternative pronunciations for acronyms were automatically generated. Further, pronunciation probabilities were automatically obtained from the *trn11* data.

2.3. Baseline models

The baseline AM used in these experiments is the one presented in [5]. It was trained on a untranscribed corpus containing 173 hours of speech data and using a lattice-based unsupervised training procedure after six incremental iterations. This model uses PLP features and is gender-independent. It covers about 15k phone contexts with 11.5k tied states and 32 Gaussians per state. Silence is modeled by a single state with 2048 Gaussians. A maximum-likelihood linear transform was applied to the model.

The baseline LM was built by interpolation of component LMs trained from all available sources, except the *trn11* transcriptions. The 4-gram baseline model gives a perplexity of 142 on the development set.

2.4. Metrics

To compare the supervised and unsupervised training methods, the "WER Recovery" metric [8] was used. It measures what fraction of the absolute gain of supervised training is recovered by the unsupervised training, i.e.:

$$\mathrm{WER}_{Rec} = \frac{\mathrm{WER}_I - \mathrm{WER}_U}{\mathrm{WER}_I - \mathrm{WER}_S}$$

where WER_I, WER_U and WER_S are the word error rates obtained with the initial, the unsupervised and the supervised systems respectively.

3. UNSUPERVISED LM TRAINING

In this work, unsupervised language modeling was applied in a straightforward manner. The baseline system was used to decode the *trn11* data. With the automatic transcriptions

[1] http://www.quaero.org

Table 1. WER of systems using the baseline, supervised ('Sup') or unsupervised ('Unsup') trained LMs. WER Recovery is given in the last column. All measures are in (%).

LM	*AM*	*WER*		WER_{Rec}
		Sup.	*Unsup.*	
baseline	baseline	33.2		-
+ trn11	baseline	32.7	33.1	20.0

obtained, a component LM was estimated. Only the best hypothesis given by the decoder was used, without any filtering or weighting technique having been applied. This model was interpolated with the 10 component LMs estimated from the remaining available sources. The highest interpolation coefficient was associated with the automatic transcriptions (> 0.2), even if they correspond to less than 0.1% of the total amount of data. This highlights the importance of transcriptions in language modeling, even if they contain errors. The perplexity of the interpolated 4-gram LM obtained was 137 on the *dev* set. For comparison, when the manual transcriptions were used instead of the automatic ones, the component LM received a coefficient of 0.3 and the perplexity of the interpolated model was reduced to 127.

A first experiment was performed in order to compare the impact of supervised and unsupervised language model training. Table 1 summarizes the speech recognition results obtained. In the first row, the baseline WER is given. When the manual or automatic transcriptions of the *trn11* set were added to the LM, absolute WER reductions of 0.5% and 0.1% were respectively obtained. The WER Recovery is therefore only 20.0% for the LM-UT in this case. This result is similar to that obtained by [9] and indicates that using automatic transcriptions in language modeling is a challenging task and needs a more extensive investigation. A possible improvement, for instance, could be take into account the confidence measures to weight the hypotheses given by the decoder.

4. UNSUPERVISED AM TRAINING

Training acoustic models is a task that requires alignment between the audio stream and its associated transcriptions. During supervised training, a forced alignment is performed using the manual transcriptions. In the unsupervised approaches, an initial system is used to decode a large amount of untranscribed data. The acoustic model parameter estimation is guided by one or many alignment hypotheses given by the decoder. In these experiments, multiple hypotheses weighted by their posterior probabilities were used, since this approach was found to lead to slightly better performances [5]. For each of the models estimated, only one iteration of unsupervised training was applied. The baseline models described in Section 2.3 were used in the initial system.

There has been an increasing use of discriminative features produced by a MLP network in speech recognition systems. It has been shown that such features lead to improvements over the state-of-the-art LVCSR systems for different languages and tasks. In this section, the results of AM-UT using PLP and MLP features are reported.

4.1. Training with PLP features

Three different PLP-based models were compared in the experiments on unsupervised acoustic training. The first model was trained only on the *trn11* set. A second model (pooled) was built using the baseline and the *trn11* data. The third model was obtained by adapting the pooled model to the *trn11* data. In fact, during preliminary tests, no difference of performance was observed when adapting the baseline or the pooled model. However, this latter was used in these experiments because of its higher likelihood. All three models have 11.5k tied states. The model trained on the *trn11* data covers about 12k phone contexts, while the other two cover 16k.

Table 2 shows the results obtained for the supervised and unsupervised systems that use PLP-based acoustic models. The first row shows the baseline system performance. Using the baseline language model and the adapted acoustic model (2nd row), WERs of 28.5% and 31.0% were obtained for the supervised and unsupervised cases, respectively. The WER Recovery is therefore 46.8% in this case. This recovery rate is much higher than what was obtained for the LM-UT (20.0%), what may explains why the unsupervised techniques have been more commonly used in acoustic modeling. However, when both unsupervised techniques were applied together (last row), the system performance was further improved, with an absolute WER reduction of 2.3% compared to the baseline system. The equivalent supervised system led to an absolute reduction of 5.1%. The three last rows present the results with systems that use the adapted language model and the three different acoustic models. The adapted acoustic model performed better than the other two. In the unsupervised case, it led to an absolute WER reduction of 0.6% compared to the model trained on the *trn11* data and 0.2% over the pooled model. In the supervised case, the WERs obtained with the model trained on the *trn11* data and the adapted model were respectively 28.9% and 28.1%.

4.2. Training with MLP features

The MLP features are extracted from a 4-layer bottleneck network [12, 13] and are generated in two steps. In the first step, the MLP network is trained using as input a raw feature vector that covers a wide temporal context (100-500 ms). This work makes use of the TRAP-DCT features [16], which are obtained from a 19-band Mel scale spectrogram, with 25 LPC coefficients for each frequency band, resulting in a 475 raw input vector. A discrete cosine transform (DCT) is applied to each band. The output layer uses the HMM phone states as target (109 phone-state targets were defined). In the second

Table 3. WER of systems built using different levels of supervision. The 'manual' and 'auto' tags indicate if the models were trained using whether the manual (supervised) or automatic (unsupervised) transcriptions of the *trn11* set. WER Recovery is given in the last column. All measures are in (%). As a reminder, the baseline WER is 33.2%

System	*MLP*	*HMM*	*LM*	*WER*	WER_{Rec}
supervised	manual	manual	+ manual	**26.9**	-
$MLP_{unsup}HMM_{sup}$	auto	manual	+ manual	27.7	87.3
$MLP_{sup}HMM_{unsup}$	manual	auto	+ manual	29.2	63.5
$MLP_{unsup}HMM_{unsup}$	auto	auto	+ manual	30.0	50.8
unsupervised	auto	auto	+ auto	**30.5**	42.9
unsupervised, pooled	auto	pooled	+ auto	**30.0**	-
unsupervised, adapted	auto	pooled $\rightarrow$ adapted	+ auto	30.1	-

Table 2. WER of baseline system and systems built using supervised ('Sup') or unsupervised ('Unsup') methods with PLP-based acoustic models. WER Recovery is given in the last column. All measures are in (%).

LM	*AM*	*WER*		WER_{Rec}
		Sup.	*Unsup.*	
baseline	baseline	33.2		-
baseline	pooled $\rightarrow$ adapted	28.5	31.0	46.8
+ trn11	trn11	28.9	31.5	39.5
+ trn11	pooled	-	31.1	-
+ trn11	pooled $\rightarrow$ adapted	**28.1**	**30.9**	45.1

step, the raw features are processed by the MLP and the features are not taken from the output layer of the MLP but from the "bottleneck" hidden layer, whose size is defined according to the desired number of features (39 in this work). Afterwards the extracted feature vector is decorrelated by a PCA transformation. The STT system thus uses a 81-parameter feature vector resulting from the concatenation of the MLP, PLP and F0 features (MLP+PLP+F0).

Using MLP features clearly increases the complexity of acoustic model training, since the parameters of the MLP network have to be estimated in addition to the HMM parameters. On unsupervised training, this new step adds another degree of uncertainty. The impact of using manual or automatic transcriptions to guide the parameter estimation for each of the main components of the system was evaluated.

Table 3 summarizes the results obtained. The upper part of the table shows systems built with different levels of supervision, where the acoustic models were trained only on the *trn11* data. As a reminder, the baseline WER is 33.2%. The fully supervised system represents the upper-bound performance level and gives an absolute WER reduction of 6.3% compared to the baseline. The system in which only the MLP network was trained in a unsupervised manner ($MLP_{unsup}HMM_{sup}$) leads to an absolute loss of performance of 0.8% in comparison with the supervised system. However, when only the HMM was unsupervised trained ($MLP_{sup}HMM_{unsup}$), this difference of performance increases to 2.3%. It may suggests that the HMM parameter estimation is more sensitive to the errors present in the automatic transcriptions used on training. In other words, the parameter estimation of the MLP network seems to be more robust to the uncertainty of the training data. When the parameter estimation of both, MLP and HMM, are derived from the automatic transcriptions ($MLP_{unsup}HMM_{unsup}$), the absolute loss of performance is 3.1% compared to the supervised system. Finally, the fully unsupervised system outperforms the baseline with an absolute WER reduction of 2.7%, representing a WER Recovery of 42.9%.

Two other unsupervised systems were tested and are shown in the lower part of Table 3. In the first one, a pooled acoustic model trained on the baseline and *trn11* data was used (*unsupervised, pooled*). In the second, this model was adapted to the *trn11* data (*unsupervised, adapted*). As expected, both models outperformed the acoustic model trained only on the *trn11* data, as was observed for PLP models. However, in the case of the MLP-based models, the pooled model performed slightly better than the adapted one, with WERs of 30.0% and 30.1%, respectively.

4.3. Comparing PLP and MLP models

In the last sections, supervised and unsupervised training of PLP and MLP based models were compared. Table 4 summarizes the results obtained using the adapted language models and acoustic models trained on the *trn11* data. The results in the same row show that models trained with MLP features outperforms the models trained only with PLP features for both, the supervised and the unsupervised systems. However, the relative improvement in the unsupervised case is lower (3.2% against 6.9%). In the same column, the supervised and unsupervised systems are compared. The relative improvement of the supervised system over the unsupervised one is 8.3% when only PLP features were used and 11.8% when they were combined with MLP features. In the last row, the WER Recovery rates are shown. The WER Recovery obtained with respect to the baseline is 39.5% for the PLP-based model and 42.9% for the MLP-based model. Although the use of MLP features increases the number of pa-

Table 4. WER of supervised and unsupervised systems built with the adapted LMs and the AMs trained only on *trn11* data using PLP+F0 or MLP+PLP+F0 features. WER Recovery is given in the last row. All measures are in (%).

System	*PLP+F0*	*MLP+PLP+F0*
unsupervised	31.5	30.5
supervised	28.9	26.9
WER_{Rec}	39.5	42.9

rameters to be estimated during acoustic training, the unsupervised MLP-based model is able to recover a more important fraction of the absolute gain obtained by the equivalent supervised model, in comparison to the PLP model.

5. CONCLUSIONS

In this paper, unsupervised training in all the components of a Portuguese broadcast system recognition was investigated. The unsupervised LM led to a slight gain of performance over the baseline. This gain was maintained when both, unsupervised acoustic and language modeling were applied. The use of MLP features in unsupervised acoustic model training was introduced. It was shown that such discriminative features help to improve the system performance even when no manual transcriptions are available. They also led to better WER Recovery rate in comparison to the PLP models.

The experiments presented in this paper were performed using only 56 hours of speech data in order to compare supervised and unsupervised training approaches. While a fully supervised system still outperforms the unsupervised one, this relative difference is on the order of 10%. It is important to note that new audio data, when available, can be easily added to the unsupervised training process. So, we expect that this difference can be reduced applying an incremental unsupervised training method. Additionally, other techniques are being investigated to improve the unsupervised language model training in order to take into account the confidence measures given by the decoder.

6. REFERENCES

[1] G. Zavaliagkos and T. Colthurst, "Utilizing untranscribed training data to improve performance," in *DARPA Broadcast News Transcription and Understanding Workshop*, February 1998, pp. 301–305.

[2] L. Wang, M. J. F. Gales, and P. C. Woodland, "Unsupervised training for mandarin broadcast news and conversation transcriptions," in *ICASSP*, Honolulu, Hawaii, April 2007, vol. IV, pp. 353–356.

[3] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: recent experiments," in *Eurospeech*, Budapest, Hungary, September 1999, pp. 2725–2728.

[4] L. Lamel and B. Vieru, "Development of a speech-to-text transcription system for finnish," in *SLTU*, Penang, Malaysia, May 2010, pp. 62–67.

[5] T. Fraga-Silva, L. Lamel, and J.-L. Gauvain, "Lattice-based unsupervised acoustic model training," in *ICASSP*, Prague, Czech Republic, May 2011, pp. 4656–4659.

[6] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, pp. 115–129, 2002.

[7] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, "Unsupervised training on large amounts of broadcast news data," in *ICASSP*, May 2006, vol. III, pp. 1056–1059.

[8] J. Ma and R. Schwartz, "Unsupervised versus supervised training of acoustic models," in *Interspeech*, September 2008, pp. 2374–2377.

[9] S. Novotney, R. Schwartz, and J. Ma, "Unsupervised acoustic and language model training with small amounts of labelled data," in *ICASSP*, Taipei, Taiwan, April 2009, pp. 4297–4300.

[10] M. Bacchiani and B. Roark, "Unsupervised language model adaptation," in *ICASSP*, Hong Kong, April 2003, vol. I, pp. 224–227.

[11] H. Hermansky, "Perceptual linear prediction (plp) analysis for speech," *Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, April 1990.

[12] P. Fousek, L. Lamel, and J.-L. Gauvain, "Transcribing broadcast data using MLP features," in *Interspeech*, Brisbane, Australia, September 2008, pp. 1433–1436.

[13] P. Fousek, L. Lamel, and J.-L. Gauvain, "On the use of MLP features for broadcast news transcription," *Text, Speech and Dialogue, Lecture Notes in Computer Science*, vol. 5246, pp. 303–310, 2008.

[14] L. Lamel, J.-L. Gauvain, V. B. Le, I. Oparin, and S. Meng, "Improved models for mandarin speech-to-text transcription," in *ICASSP*, Prague, Czech Republic, May 2011, pp. 4660–4663.

[15] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, pp. 89–108, 2002.

[16] P. Schwarz, P. Matjka, and J. Cernocky, "Towards lower error rates in phoneme recognition," in *International Conference on Text, Speech and Dialogue*, Brno, Czech Republic, September 2004, pp. 465–472.