# The Vocapia Research ASR Systems for Evalita 2011

Julien Despres[1], Lori Lamel[1,2], Jean-Luc Gauvain[1,2], Bianca Vieru[1], Cécile Woehrling[1], Viet Bac Le[1], and Ilya Oparin[2]

[1]Vocapia Research, 3 rue Jean Rostand, 91400 Orsay, France
[2]CNRS-LIMSI, 91403 Orsay, France
{despres,lamel,gauvain,vieru,woehrling,levb}@vocapia.com,oparin@limsi.fr
http://www.vocapia.com http://www.limsi.fr/tlp

**Abstract.** This document describes the speech recognizers submitted by Vocapia Research to the Evalita 2011 evaluation for the open unconstrained automatic speech recognition (ASR) task. The aim of this evaluation was to perform automatic speech recognition of parliament audio sessions in the Italian language. Two systems were submitted. The primary system has a single decoding pass and was optimized to run in real time. The contrastive system, developed in collaboration with LIMSI-CNRS, has two decoding passes and runs in about 5×RT. The case-insensitive word error rates (WER) of these systems on the Evalita development data are respectively 10.2% and 9.3%.

**Keywords:** automatic speech recognition, speech-to-text transcription, Italian, Evalita, unsupervised adaptation, MLP features, Neural Network language models, PLP, MMIE, SAT, MAP.

## 1 Introduction

Vocapia Research took part in the first Evalita evaluation campaign on large vocabulary automatic speech recognition (ASR) for the Italian language. An Italian system had first been developed using internal corpora. It was then improved in the context of the Quaero program[1], where participants benefit from shared training and development corpus, and for which periodic evaluations are organized. For the Evalita evaluation, Vocapia updated and adapted the existing ASR Italian system to the parliament session transcription task.

This paper describes the specific work done in preparation for our participation in the 2011 Evalita evaluation. A description of the system will be given including the text normalization procedure, and the language models, acoustic models and pronunciation lexicon. The results obtained on the Evalita development data are provided, illustrating the improvements obtained by incorporating the Evalita training corpora in the language and acoustic models.

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

## 2 System Description

Two systems were submitted to the EVALITA 2011 evaluation campaign: one single pass system running in real time and a two pass system developed in collaboration with LIMSI-CNRS, running in about 5×RT.

### 2.1 Text Normalization and Language Models

The case of the training corpus was checked and corrected using a specific 3-gram language model. A lattice was generated from the training corpus letting the language models choose between the different forms observed in the whole corpus. For example, the word "sky" appears as "SKY", "SKy", "SkY", "Sky", or "sky" and the 3-gram language model will have to select the right case depending on the context. As the main problem with the case concerns the first word of each sentence, it can be considered that the word caseing inside a sentence is generally correct. So the original word caseing is always privileged by a higher weight in the lattice.

The decoding parameters were optimized on an internal development data set so as to obtain a good compromise between generated and corrected errors. Finally the first letter of the first word of each sentence was uppercased. The best setup reduces the difference between the case-sensitive and case-insenstive word error rates to 0.3% absolute

A 100k word vocabulary was used in order to maximize the lexical coverage (1.34% of Out Of Vocabulary words) while maintaining a system able to decode in real time.

The language models were trained on a corpus of 960 million words listed in Table 1. Component 4-gram LMs were trained on each subcorpus, and then interpolated to form the final LM. The interpolation coefficients were automatically computed so as to minimize the perplexity on the EVALITA development data set.

**Table 1.** Description of the Italian training text corpus. The number of words is computed after the normalization.

| **Italian sources** | **Epoch** | **Words** | interpolation coef. |
|---|---|---|---|
| La Stampa | 1992-2000 | 226M | 0.10 |
| Google news | 2008-2011 | 429M | 0.05 |
| manual transcriptions | 1992-2011 | 1M | 0.09 |
| Various web data | 1997-2007 | 122M | 0.11 |
| Various web data | 2008-2010 | 141M | 0.18 |
| Evalita data | - | 29M | 0.44 |
| EPPS | - | 7M | 0.30 |
| Total | 1992 to 2011 | 958M | 1.00 |

Neural network LMs (NNLMs) were used for final lattice rescoring in the two pass system. In contrast to conventional N-gram LMs in which words are represented in a discrete space, Neural network LMs (NNLMs) make use of continuous-space representation of words, which enables a better estimation of unseen N-grams. The neural network deals with two tasks: projection of words with history to continuous space and calculation of LM probabilities for the given history. NNLMs have been shown to improve over the N-gram baseline for different languages and tasks [9].

Four different neural networks were generated with different number of nodes in the hidden layer. The networks vary in the size of the hidden layer (500, 450, 500, 430), and the projection size of P-dimensional continuous space (300, 250, 200, 220). Three previous words form an input to the NN, and the 12k most frequent words are used as a shortlist to estimate the probabilities at the output layer as described in [10],[9]. Since it is not feasible to train a NNLM on all the available texts, the data used to train the NNLMs was selected according to the interpolation weights of the component N-gram LMs in the baseline N-gram LM. Only the top four corpora according to N-gram LM interpolation weights were used to train the NNLMs. With the NNLM the perplexity of the EVALITA development data is reduced from 162 with the interpolated N-gram LM to 142.

### 2.2 Acoustic Models

The acoustic features used are a concatenation of PLP-like [6] with probabilistic features produced by Multi Layer Perceptron (MLP) [1]. As in [3], 39 cepstral parameters are derived from a Mel frequency spectrum, with Cepstral mean removal and variance normalization carried out on a segment-cluster basis, resulting in a zero mean and unity variance for each cepstral coefficient. TRAP-DCT features are obtained from a 19-band Bark scale spectrogram, using a 30 ms window and a 10 ms offset. A discrete cosine transform (DCT) is applied to each band (the first 25 DCT coefficients are kept) resulting in 475 raw features, features which are the input to a 4-layer MLP with the bottleneck architecture [5]. The size of the third layer (the bottleneck) is equal to the desired number of features (39). A 3-dimensional pitch feature vector (pitch, $\Delta$ and $\Delta\Delta$ pitch) is combined with the other features, resulting in a total of 81 parameters (MLP+PLP+f0).

A MLP network was trained for Italian using the simplified training scheme proposed in [11] on about 87 hours of data from a variety of broadcast sources. The training data are randomized and split in three non-overlapping subsets, used in 6 training epochs with fixed learning rates.

The first 3 epochs use only 13% of data, the next 2 use 26%, the last epoch uses 52% of the data, with the remainder used for cross-validation to monitor performance. The MLP has 84 targets, corresponding to the individual states for each phone and one state for each of the additional pseudo phones (silence, breath, filler).

As in [3] the acoustic models are tied-state, left-to-right 3-state HMMs with Gaussian mixture observation densities (typically 32 components). The triphone-

based phone models are word independent, but position-dependent. The states are tied by means of a decision tree to reduce model size and increase triphone coverage. The acoustic models are gender-dependent and speaker adaptive trained (SAT). Silence is modeled by a single state with 1024 Gaussians.

The AM were trained in an unsupervised manner [7] on about 120h of detailed manual transcriptions mainly from previous European or national projects plus 30h of audio data distributed in EVALITA. These acoustic models cover 8k phone contexts. They are discriminatively MMI trained and use probabilistic features based on bottleneck multi-layer perceptrons (MLP) and modified TRAP-DCT features. Combined with classical PLP features, these probabilistic features significantly reduce the word error rate. A maximum a posteriori (MAP) [4] adaptation to the EVALITA training corpus was made using the automatic transcriptions produced by our baseline system.

### 2.3 Lexicon and Phone Set

The pronunciations of the words in the vocabulary were automatically generated with a rule-based phoneticizer using the set of 30 phones given in Table 2). No specific phones were used for geminated consonants, so the phoneticizer simply doubles these consonants in the generated pronuncations. Similarly, affricates are treated as sequences of phones: /tʃ/, /ts/ and /dz/ (geminated forms /ttʃ/, /tts/ and /ddz/).

**Table 2.** The VR phone set used to represent pronunciations the Italian lexicon.

| VR phone | IPA | Example | VR phone | IPA | Example |
|---|---|---|---|---|---|
| Consonants | | | Vowels | | |
| ç | ʃ | sciamano | a | a | pane |
| z | z | caso | e | e | era |
| s | s | secondo | i | i | primo |
| k | k | tragico | o | o | sano |
| j | d͡ʒ | dirigere | u | u | ultima |
| g | g | spiegare | è | ɛ | è |
| ñ | ɲ | gnocco | ò | ɔ | peró |
| ý | ʎ | gli | w | w | guerra |
| l | l | solo | y | j | spiega |
| p | p | tempo | Non-speech symbols | | |
| b | b | problema | . | silence | |
| t | t | cento | | [breath] | |
| d | d | edizione | & | [fw] | |
| f | f | conferenza | | | |
| v | v | nove | | | |
| m | m | come | | | |
| n | n | fine | | | |
| r | r | ancora | | | |

### 2.4 Decoding Results

The baseline system was optimized to decode broadcast news (BN) data. The adaptation of the baseline system (acoustic and language model) to the parliament session domain improved the word error rate by 1.9% absolute. It appeared that the adapted system decodes parliament data much faster than BN data. The system was slowed down so as to run in real time on the EVALITA development data set.

The first step in processing an audio document is to segment and partition the data, identify the portions containing speech data to be transcribed [2] and associating segment cluster labels, where each segment cluster ideally represents one speaker.

The primary submitted system decodes in a single pass and runs in 1×RT on two cores of an Intel i5-2500 processor. It achieves a word error rate of 10.2% on the EVALITA development data set as shown in Table 3. An absolute gain of 3.0% was obtained between the primary system submitted and our initial baseline system.

In the contrastive system word decoding is carried out in two decoding passes. Each decoding pass produces a word lattice with cross-word, word-position dependent acoustic models, followed by consensus decoding with a 4-gram language model and pronunciation probabilities. Unsupervised acoustic model adaptation is performed for each segment cluster using the CMLLR and MLLR [8], and the lattices produced are rescored by the neural network LM interpolated with a 4-gram back-off LM. The contrastive two pass system runs in about 5×RT and obtains a word error rate of 9.3% on the development data.

**Table 3.** Word error rates, out-of-vocabulary rate and perplexity of the VR Italian transcription system computed on the EVALITA development data set. The baseline system was optimized for broadcast news data.

| System | voc. size | WER (%) | OOV (%) | ppl 4-g | ×RT |
|---|---|---|---|---|---|
| Baseline | 65k | 13.2 | 2.0 | 218 | 0.3 |
| + add Evalita data to LM | 100k | 11.9 | 1.1 | 162 | 0.3 |
| + add Evalita data to AM | 100k | 11.3 | 1.1 | 162 | 0.3 |
| + slower decode | 100k | 10.2 | 1.1 | 162 | 1.0 |
| 2 pass system | 100k | 9.3 | 1.1 | 142 | 5.0 |

### 2.5 Conclusions

This paper has described the speech transcription systems used for the Vocapia Research submissions to the EVALITA 2011 evaluation for the open unconstrained automatic speech recognition (ASR) task. The paper highlighted the specific work done in preparation for this participation, including the normalization of

texts for language model training, the generation of the pronunciation dictionary, and adapting the acoustic and language models to the parliamentary task. The results on the EVALITA development data show the improvement from an initial word error rate of 13.2% to 10.2% for a real-time system, and to 9.3% for a 2-pass contrastive system developed in collaboration with LIMSI-CNRS. On the evaluation data these systems obtained word error rates of 6.4% and 5.4%, respectively.

## Acknowledgment

This work was partly realized as part of the Quaero Project, funded by OSEO, the French State agency for innovation.

## References

1. P. Fousek, L. Lamel and J.-L. Gauvain, "On the Use of MLP Features for Broadcast News Transcription", *TSD'08*. LNCS 5246/2008, 303.10, Springer Verlag, Berlin/Heidelberg (2008)
2. J.L. Gauvain, L. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," *ICSLP'88*, Sydney, Australia, pp. 1335-1338, December 1998
3. J.-L. Gauvain, L. Lamel and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication,* 37(1-2):89-108, (2002)
4. J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994. [Online]. Available: `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.6428`
5. F. Grézl and P. Fousek, "Optimizing Bottle-Neck Features for LVCSR", *IEEE ICASSP'08*, pp. 4729-4732, Las Vegas, (2008)
6. H. Hermansky, "Perceptual linear prediction (plp) analysis for speech," *J. Acoust. Soc. Amer.*, 87:1738-1752, (1990)
7. 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)
8. C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, **9**(2), pp. 171-185, (1995)
9. H. Schwenk and J.L. Gauvain, Training Neural Network Language Models On Very Large Corpora, *JHLT/EMNLP*, pp. 201-208, Vancouver, (2005)
10. H. Schwenk, *Continuous Space Language Models*, Computer, Speech & Language, **21**:492-518, (2007)
11. Q. Zhu, A. Stolcke, B.Y. Chen and N. Morgan, "Using MLP features in SRI's conversational speech recognition system", *Interspeech'05*, pp. 2141-2144, Lisbon, Portugal, September 2005