THE BAVIECA OPEN-SOURCE SPEECH RECOGNITION SYSTEM EXPERIMENTS ON ADULT AND CHILDREN ITALIAN SPEECH

Piero Cosi*, Giulio Paci*, Giacomo Sommavilla*, Fabio Tesser* and Daniel Bolaños**
*Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche
Unità Organizzativa di Supporto di Padova - Italy
[piero.cosi, giulio.paci, giacomo.sommavilla, fabio.tesser]@pd.istc.cnr.it

** Boulder Language Technologies
Boulder, CO, USA 80301
dani@bltek.com

1. ABSTRACT

In this paper, we describe the adaptation of the new BAVIECA ASR engine to Italian. After a brief comparison with other similar systems, we give an overview of the BAVIECA Acoustic and Language Model (AM) training procedure, we describe the final test procedures and we end reporting the results obtained so far in some adults and children speech ASR experiments, comparing them with those previously obtained with different recognition engines.

2. INTRODUCTION

During the last few years, many different Automatic Speech Recognition frameworks have been developed for research purposes and, nowadays, various open-source automatic speech recognition (ASR) toolkits are available to research laboratories. Systems such as HTK (Young et alii, 2009), SONIC (Pellom, 2001), (Pellom and Hacioglu 2003), SPHINX (Lee et alii, 1990), (Walker et alii, 2004), RWTH (Rybach et alii, 2009), JULIUS (Lee et alii, 2001), KALDI (Povey and Ghoshal, 2009), the more recent ASR framework SIMON (SIMON, 2014), and the relatively new system called BAVIECA (BAVIECA, 2014), (Bolaños, 2012), are a simple and probably the most famous list.

The experience that we had with some of these systems at ISTC-CNR of Padova convinced us in the past that CSLR SONIC and CMU SPHINX are two versatile and powerful tools to build a state-of-the-art ASR. Thus, we decided to benchmark them in different contexts (Cosi et alii, 2007), (Cosi & Nicolao, 2009). In our former work, we performed a recognition test on continuous clean speech for Arabic language (Cosi et alii, 2007) with a 1k-word vocabulary. Although the task was quite simple, the result was encouraging (in terms of Word Error Rate (WER). SONIC scored 1.9% and SPHINX did 1.3%). The simplicity, with which we were able to configure the systems for such a phonetic and spelling-complicated language, was the most interesting feature of these experiments. In our more recent work, these two systems were tested in an evaluation campaign on the automatic recognition of connected digit for Italian language, named EVALITA (Cosi & Nicolao, 2009). The results were also extremely good (the SONIC word error rate was 2.7% and the SPHINX one was 4.5%) and both systems yielded the best performances among the other competitors.

From all these considerations, this paper aims at making a useful comparison between the new BAVIECA engine and some of the best open source ASR systems, and at demonstrating its effectiveness even on a difficult target task, such as the application of ASR sys-

tems to children speech. Therefore, the present work focuses on applying the BAVIECA toolkit to the same children's speech recognition task where SPHINX and SONIC had good results in our past experiences (Cosi & Pellom, 2005), (Cosi, 2008), (Cosi, 2009).

3. BAVIECA

The BAVIECA toolkit includes a set of command line tools that can be used to build very sophisticated large vocabulary speech recognition systems from scratch. According to (Bolaños, 2012), "BAVIECA is an open-source speech recognition toolkit intended for speech research and system development. The toolkit supports lattice-based discriminative training, wide phonetic-context, efficient acoustic scoring, large n-gram language models, and the most common feature and model transformations. BAVIECA is written entirely in C++ and presents a simple and modular design with an emphasis on scalability and reusability. BAVIECA achieves competitive results in standard benchmarks. The toolkit is distributed under the highly unrestricted Apache 2.0 license, and is freely available on SourceForge".

Moreover, as written in his official web site (http://www.bavieca.org/tools.html#), BAVIECA "offers an Application Programming Interface (API) that exposes speech processing features such as speech recognition, speech activity detection, forced alignment, etc. This API is provided as a C++ library that can be used to create stand-alone applications that exploit BAVIECA's speech recognition features".

Compared to existing open-source automatic speech recognition (ASR) toolkits, such as HTK (Young et alii, 2009), CMU-Sphinx (Lee et alii, 1990), (Walker et alii, 2004), RWTH (Rybach et alii, 2009), JULIUS (Lee et alii, 2001) and the more recent Kaldi (Povey-Ghoshal, 2009), BAVIECA is characterized by a simple and modular design that favors scalability and reusability, a small code base, a focus on real-time performance and a highly unrestricted license.

As illustrated in the BAVIECA web page (www.bavieca.org) the list below summarizes the main features of the BAVIECA toolkit.

- Large vocabulary continuous speech recognition
  - Dynamic search decoder with support for cross-word triphone and pentaphone HMMs
  - Weighted Finite State Acceptor (WFSA) based speech decoder and efficient WFSA network builder (cross-word triphones)
  - Efficient computation of emission probabilities thanks to the use of the nearest neighbor approximation, partial distance elimination and support for Single Instruction Multiple Data (SIMD) parallel computation (x86 architecture only)
  - Lattice generation (both decoders)
  - Hypothesis files in NIST formats (SCLITE can be use for scoring hypotheses)
- Acoustic modeling
  - Acoustic models based on continuous density Hidden Markov Models (CD-HMMs) with emission probabilities modeled using mixtures of Gaussian distributions (GMMs)
  - HMM topology fixed to three states left to right
  - Variable number of Gaussian components per HMM-state

    - No explicit modeling of transition probabilities
    - Diagonal and full covariance modeling
    - Cross-word context dependency modeling using triphone, pentaphones, heptaphones, etc
    - Maximum Likelihood Estimation criterion
    - Discriminative training using boosted Maximum Mutual Information (bMMI) criterion with I-smoothing and cancellation of statistics
    - Parallel accumulation of sufficient statistics for both Maximum Likelihood and Discriminative Training criteria
    - Linear algebra support through template classes (Matrix, Vector, etc) wrapping third party libraries (BLAS and LAPACK)
- Language modeling
    - Support for n-gram language models in ARPA and binary formats
    - Support for any n-gram order (zerogram, unigram, bigram, trigram, fourgram, etc)
    - Language models are internally represented as Finite State Machines
- Speaker adaptation
    - Model space Maximum Likelihood Linear Regression (MLLR) using regression trees to automatically determine the number of transforms to be used and how adaptation data is shared among transforms
    - Feature space Maximum Likelihood Linear Regression (fMLLR)
    - Vocal Tract Length Normalization (VTLN)
- Feature extraction
    - Mel Frequency Cepstral Coefficients (MFCC) features
    - Cesptral Mean Normalization (CMN) and Cepstral Mean Variance Normalization (CMVN) at both utterance or session level
    - Feature decorrelation and dimensionality reduction using Heteroscedastic Linear Discriminant Analysis (HLDA)
    - Support for spliced features and third order derivatives
- Lattice processing and n-best list generation
    - Lattice rescoring using different criteria: maximum likelihood or posterior probabilities
    - N-best generation (from lattices) using different criteria: maximum likelihood or posterior probabilities
    - Lattice word error rate (WER) computation (oracle)
    - Lattice alignment and HMM-state marking
    - Attach LM-scores to lattice edges according to a given language model
    - Lattice-based posterior probability computation
    - Confidence annotation
    - Lattice path-insertion (discriminative training)
    - Lattices are processed in binary format but text format is available for readability purposes
- Speech activity detection
    - HMM-based speech activity detection

4. EXPERIMENTAL FRAMEWORK

Both Italian adult and children speech recognition were considered. Especially the second one is a very peculiar task and its difficulties have been investigated several times by the ISTC-CNR research group (Cosi & Pellom, 2005), (Cosi, 2008), (Cosi, 2009).

4.1.Dataset

In the experiment here described, two Italian Corpora have been tested:

- Italian FBK APASCI (Angelini et alii, 1993); APASCI is an Italian speech database recorded in a silent room with a Sennheiser MKH 416 T microphone; it includes 5,290 phonetically rich sentences and 10,800 isolated digits, for a total of 58,924 word occurrences (2,191 different words) and 641 minutes of speech material which was read by 100 Italian speakers (50 male and 50 female).
- Italian FBK ChildIt Corpus (Gerosa et alii, 2007); this is a corpus of Italian children voice that contains almost 10 hours of speech from 171 children (85 females and 86 males) aged between 7 and 13 (from grade 2 up to grade 8), who were all native speakers from regions in the north of Italy. Each child provided approximately 50-60 read sentences extracted from age-appropriate literature. The audio was sampled at 16 kHz, 16 bit linear, using a Shure SM10A head-worn microphone.

According to (Cosi & Pellom, 2005), to test the system, the corpus was divided into a training set consisting of 129 speakers (64 females and 65 males) and a test set consisting of 42 speakers (21 females and 21 males) balanced by gender and aged between 7 and 13. Training and test sentences that contain mispronunciation and noisy words have been excluded from the following experiments, while all other sentences, even those with annotated extra-linguistic phenomena, such as human disturbs (lip smacks, breath, laugh, cough, etc.), generic noises non-overlapping with speech have been included. The orthographic transcriptions of the prompt sentences have been used for the training and the automatic phonetic transcription have been used for the test. Actually, slightly mispronounced words, such as the interrupted ones, could still have been considered by modifying their phonetization and forcing a silence tag at the end of them. This helped to prevent co-articulation between the interrupted token and the following word that could lead to improper model training. This was also justified by the human speech error explanation theories, which affirm that, after the occurrence of an error in speech production, there is always a little pause (~20 ms) due to the time spent for the interrupting and the restarting actions.

4.2. Training and Decoding Process

As most of the ASR systems require, a list of words with their standard phonetization (lexicon), a list of extra-text words (fillers), a list of phonemes, a list of questions for tree-clustering and the accurate transcription of each training audio file must be provided to configure the system correctly. Phone list and decision-tree question structure, both compiled by expert Italian linguists, are the same as those used in our previous work (Cosi & Hosom, 2000). Differently from APASCI, only orthographic transcriptions were available within the FBK-ChildIt corpus, thus, an automatic phonetic alignment had to be provided. In our previous works on children speech with SONIC (Cosi & Pellom, 2005), (Cosi, 2008), (Co-

si, 2009), this has been automatically obtained by aligning text to the audio data with the Italian AM previously trained on APASCI adult speech (Cosi & Hosom, 2000), while in this work the phonetic alignment was obtained with a raw AM trained by uniformly segmenting audio data. The latter method yields to a less accurate bootstrap, but it allows for flexibility in choosing transcriptions and phoneme list. The main characteristics of the training and decoding processes are summarized in Figure 1.

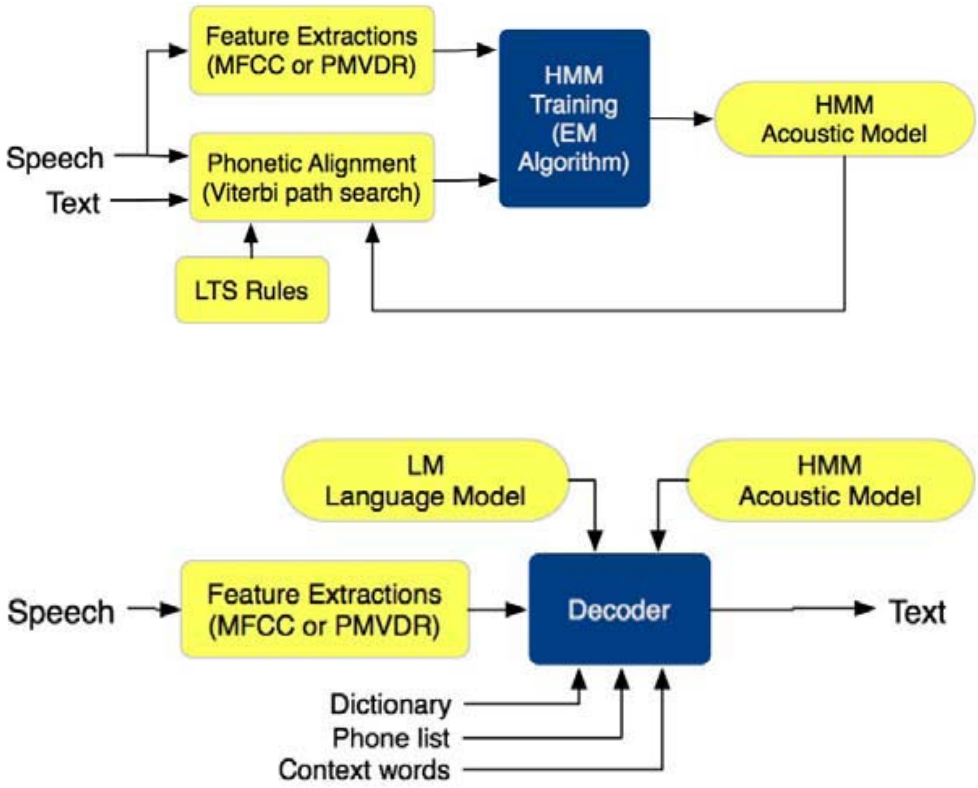

Figure 1. General scheme of ASR training (upper) and decoding (lower) processes. In the upper part, LTS means “letter to sound”, and EM means “expectation maximization”.

5. RESULTS

The experiments have been carried out with and without MLLR (Maximum Likelihood Linear Regression) adaptation, and in all cases, training and test materials have been kept separate. The results, shown in Table 1, are quite encouraging and comparable to those obtained with similar open-source systems on standard benchmarks (Nicolao, Cosi, 2011), (Cosi e alii, 2014).

APASCI

		WCR	subs	inser	del	**WER**	SER	(%)
1-GRAM full LM	SI	85.0	9.9	5.1	0.4	**15.5**	67.0	
	MLLR	87.3	8.7	5.0	0.3	**13.8**	66.5	
3-GRAM full LM	SI	99.0	0.5	0.5	0.1	**1.1**	7.4	
	MLLR	99.0	0.5	0.5	0.1	**1.1**	7.0	

ChildIt

		WCR	subs	inser	del	**WER**	SER	(%)
1-GRAM full LM	SI	78.7	15.0	6.3	1.0	**22.3**	70.6	
	MLLR	80.3	13.7	6.1	1.0	**20.8**	67.9	
3-GRAM full LM	SI	98.3	0.9	0.7	0.2	**1.9**	8.7	
	MLLR	98.5	0.8	0.7	0.2	**1.7**	8.1	
		PCR	subs	inser	del	**PER**	SER	(%)
3-GRAM-PHN full LM	SI	83.1	8.8	8.1	0.5	**17.4**	97.2	
	MLLR	85.2	72.0	76.0	0.4	**15.2**	95.9	

Table 1: Results, expressed in terms of word and phoneme correct rate (WCR, PCR), substitutions (subs), insertions (inser), deletions (del), word and phoneme error rate (WER, PER) and sentence error rate (SER), obtained in the experiments executed on APASCI and ChildIt corpora, with MLLR (Maximum Likelihood Linear Regression) and without MLLR adaptation (SI).

Word error rate (WER) was considered both for APASCI and ChildIt, while, as illustrated in the Table, only for the ChildIt case the Phonetic Error Rate (PER) was taken in consideration for computing the score of the recognition process. As for the word case, the PER is defined as the sum of the deletion, substitution and insertion percentage of phonemes in the ASR outcome text with respect to a reference transcription. Ideally, a hand-labelled reference would have been preferred, because it would have been corrected at the phonetic level to take into account children's speech pronunciation mistakes. Since this was not available, the automatic phonetic sequences obtained by a Viterbi alignment of the word-level orthographic transcription have been used. The reference test transcriptions were created using the SONIC aligner with the general-purpose Italian model created in (Cosi & Hosom, 2000). This method was chosen because it allowed for automatically se-

lecting the best pronunciation for each word in the training data among the alternative choices available in the 400,000-word Italian lexicon available.

6. CONCLUSIONS

In conclusion, in spite of the fact that the development of BAVIECA is still a work in progress, it exhibits competitive results on standard benchmarks (Nicolao, Cosi, 2011) and it could be surely used on research projects addressing both read and conversational children's speech, which is an inherently difficult task as well as conversational adult's speech.

ACKNOWLEDGMENTS

This work was partially supported by the EU FP7 "ALIZ-E" project (grant number 248116).

REFERENCES

Angelini B., Brugnara, F., Falavigna, D., Giuliani, D., Gretter, R. & Omologo, M. (1993), "A baseline of a speaker independent continuous speech recognizer of Italian", in Proceedings of the 3rd European Conference on Speech Communication and Technology (EUROSPEECH 1993), Berlin, Germany, September 22-25, 847-850.

BAVIECA (2104). WEB site: http://www.bavieca.org.

Bolaños D. (2012), "The Bavieca Open-Source Speech Recognition Toolkit". In Proceedings of IEEE Workshop on Spoken Language Technology (SLT), December 2-5, 2012, Miami, FL, USA, 2012.

Cosi, P. (2008), "Recent Advances in Sonic Italian Children's Speech Recognition for Interactive Literacy Tutors", in Proc. of 1st Workshop On Child, Computer and Interaction (WOCCI), Chania, Greece, 2008.

Cosi, P. (2009), "On the Development of Matched and Mismatched Italian Children's Speech Recognition Systems", in Proceedings of INTERSPEECH 2009, Brighton, UK, 540-543, 2009.

Cosi, P., Hosom, J.P. (2000), "High Performance General Purpose Phonetic Recognition for Italian", in Proceedings of ICSLP 2000, Beijing, 527-530, 2000.

Cosi P., Pellom B. (2005), "Italian Children's Speech Recognition For Advanced Interactive Literacy Tutors", in Proceedings of INTERSPEECH 2005, Lisbon, Portugal, 2201-2204, 2005.

Cosi, P., Nicolao, M., Sommavilla, G. & Tisato, G. (2007), "Sviluppo di un sistema di riconoscimento per l'Arabo: problemi e soluzioni", in Proceedings of AISV 2007, 4th Conference of AISV, Reggio Calabria, Italy.

Cosi, P., Nicolao, M. (2009), "Connected Digits Recognition Task ISTC-CNR Comparison of Open Source Tools", in Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, Reggio Emilia, Italy.

Cosi, P., Nicolao, M., Paci G., Sommavilla, G. Tesser F. (2014), "Comparing Open Source ASR Toolkits on Italian Children Speech", submitted to INTERSPEECH 2014, Singapore.

Gerosa, M., Giuliani, D., and Brugnara, F. (2007). “Acoustic variability and automatic recognition of children's speech". In Speech Communication 49 (2007), 847-860.

Lee K.F., Hon H.W., and Reddy R. (1990), “An overview of the SPHINX speech recognition system". In IEEE Transactions on Acoustics, Speech and Signal Processing 38.1 (1990), 35-45.

Lee A., Kawahara T., and Shikano K. (2001). “JULIUS - an open source real-time large vocabulary recognition engine". In Proceedings of INTERSPEECH 2001, 1691-1694.

Nicolao M., Cosi P. (2011), “Comparing SPHINX vs. SONIC Italian Children Speech Recognition Systems”, in Abstract Book & CD-Rom Proceedings of AISV 2011, 7th Conference of Associazione Italiana di Scienze della Voce, “Contesto comunicativo e variabilità nella produzione e percezione della lingua”, 26-28 gennaio 2010, Università del Salento – Lecce – Abs: 85 - (CD: 414-425).

Pellom B. (2001). “SONIC: The University of Colorado Continuous Speech Recognizer”, Technical Report TR-CSLR-2001-01, Center for Spoken Language Research, University of Colorado, USA, 2001.

Pellom B., and Hacioglu K. (2003). “Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task”, Proc. ICASSP, Hong Kong, 2003.

Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlíček P., Qian Y., Schwarz P., Silovský J., Stemmer G., Veselý K. (2011), “The Kaldi Speech Recognition Toolkit,” in Proc. of ASRU, 2011.

Rybach D., Gollan C., Heigold G., Hoffmeister B., Lööf J., Schülter R., and Ney H. (2009), “The RWTH Aachen University Open Source Speech Recognition System,” in Proc. of INTERSPEECH, 2009, 2111–2114, 2009.

SIMON (2014). WEB site: http://www.simon-listens.com.

Walker W., Lamere P., Kwok P., Raj B., Singh R., Gouvea E., Wolf P., and Woelfel J. (2004), “Sphinx-4: A Flexible Open Source Framework for Speech Recognition,” Sun Microsystems Inc., Technical Report SML1 TR2004-0811, 2004.

Young S., Evermann G., Gales M., Hain T., Kershaw D., Liu X., Moore G., Odell J., Ollason D., Povey D., Valtchev V., and Woodland P. (2009), The HTK Book (for version 3.4). Cambridge Univ. Eng. Dept., 2009.