# Description of the ESAT speech recognition system - January 2006

ESAT's speech recogniser is an efficient and flexible tool that was developed during the past 15 years with the aim of speaker independent recognition of continuous speech with a large vocabulary. The system combines many of the recent advancements in automatic speech recognition with a very efficient decoder in a proven HMM architecture.

## 1. Properties of the current system

- It implements the modern probabilistic speech recognition framework completely without any compromises w.r.t. correctness of the evaluation and use of all components.
- It is very efficient w.r.t. both memory use and computational load. This is achieved by means of specific algorithms (fast acoustic model evaluation, LM-caching, LM-forwarding, transducers, …) and a highly optimised implementation.
- It is flexible: its modularity allows for research into different aspects of the recognition (preprocessing, acoustic modelling, language models, pronunciation variation, …)
- Historically, the ESAT recogniser has been used mainly for speaker independent tasks, at which it excels. Recently, several modules for off-line (MAP, MLLR) and fast on-line (VTLN) speaker adaptation have been added as well.
- It is optimised towards large vocabulary continuous speech recognition. The package can also be used for small and medium vocabulary ASR tasks. However, the present software does not provide dedicated optimisation for small vocabulary tasks, which makes it somewhat less efficient compared to embedded systems.
- It is available on a wide variety of platforms: the development platform is Linux+gcc, but the realtime recognition engine also works on several other Unixes (Mac OS X, SPARC-Solaris, HP-UX, Alpha-OSF1) and Windows without much extra effort because the code was written with an eye for compatibility between platforms and compilers. Small efforts may be needed for platform specific interfaces (e.g. for audio input). The system is available both as a library and as a set of command line tools.

## 2. Core modules of the current system

### Preprocessing

- Flexible preprocessing with a user configurable pipe-line of simple processing blocks which can be used both off-line (e.g. file to file) and on-line (front end of any program that processes a feature stream).
- Processing blocks available for all common front-end operations (FFT, mean-normalisation, Mel-filterbanks, time-derivatives, …)
- Advanced processing blocks such as: vocal tract length normalisation [1, 15], (non linear) spectral subtraction [35, 3], noise masking [27, 2], model based feature enhancement [26, 25], missing data processing [29, 30], pitch and formant detection/tracking, silence/speech detection [28], …
- New processing blocks can be added easily.
- Advanced techniques for feature selection (mutual information based linear discriminant analysis) and feature conditioning (least squares feature decorrelation) [8, 7, 4, 20].

### Acoustic modelling

States are typically modelled as a weighted sum of gaussians, but discrete models containing VQ in the preprocessing or neural networks are also available. State coupling is done with a phonetic decision tree [18, 16]. Gaussian tying is also fully supported, ranging from full tying over occupancy-based tying and phonetic tying to no tying at all [5, 19, 18, 16, 21]. The FRoG algorithm achieves a fast evaluation of the gaussian mixtures [5, 18, 4].
General scripts for initialisation and training of context independent as well as context dependent models are available. Training based on Viterbi and on existing alignments is available as a basic tool, as well as a tool to generate automatic alignments given the acoustic model (which can be used to develop corpus annotations) [11, 24, 10, 13]. Tools to generate other automatic annotations can be built easily, e.g. prosodic annotations [17].

### Language modelling

A specific protocol is used for communication between the search algorithm and the language model,

allowing the use of different types of language models. Currently N-grams and finite state grammars are implemented. Software is available to generate N-grams (Katz and Kneser-Ney backoff, Good-Turing and absolute discounting), but any existing software package can be used provided it delivers N-grams in ARPA format. A PLCG (Probabilistic Left Corner Grammar) based language model has also been implemented [31, 33, 32, 34].

**Pronunciation modelling**

All pronunciation information (pronunciation variations, cross-word context dependent phone models, assimilation rules, …) is pre-compiled in an efficient search network (cf. transducers) [6, 9, 4].

**Search algorithm**

The single pass time synchronous beam search algorithm combines the pronunciation information and the language model dynamically. This configuration allows the use of complex pronunciation models (see pronunciation modelling above) and language models with arbitrary complex contents (such as N grams with arbitrary values of N) in a fast and memory efficient manner [6, 9, 4, 14].
The result is one sentence (with the highest score) or a graph, and also some byproducts such as an alignment, acoustic scores and confidence measures [23, 22]. Several post processing steps can be applied on the generated graphs.

**(Word) lattice modules**

Flexible post-processing of the (word) lattices with a user configurable flow-chart of simple processing blocks which can be used both off-line (e.g. file to file) and on-line (back end of any program that generates lattices). This post-processing includes some very complex rescoring blocks, resulting in a very efficient multi layer recogniser [12].

**3. Benchmark results**

Research on recognition has been mainly conducted for Dutch and English, for which models are available (e.g. Dutch acoustic models based on selected components of the CGN corpus). Some benchmark results are:

- WSJ 5k closed vocabulary, speaker independent, WSJ-0 acoustic training data, bigram LM: 4.9% WER on the nov92 test set (this is our AURORA-4 clean speech reference). When using more acoustic data (WSJ-1) and a trigram, a WER of 1.8% can be achieved.
- WSJ 20k open vocabulary (1.9% OOV rate), speaker independent, WSJ-1 acoustic training data, trigram: 7.0% WER on the nov92 test set. This task runs in real time.
- Switchboard spontaneous telephone conversations, 310 hours of acoustic training data, 3M words for training of the LM, open vocabulary: 29% WER on the 2001 HUB5 benchmark. This result was obtained with a single decoding pass and without any speaker adaptation, running at 4xRT on a pentium4 2.8GHz. A similar real-time configuration gives a WER of 31%.

**4. References**

[1] Tom Claes, Ioannis Dologlou, Louis ten Bosch, and Dirk Van Compernolle. A novel feature transformation for vocal tract length normalisation in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 6(6):549–557, November 1998.
[2] Tom Claes and Dirk Van Compernolle. SNR-normalisation for robust speech recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 331–334, Atlanta, Georgia, U.S.A., May 7–10 1996.
[3] Tom Claes, Fei Xie, and Dirk Van Compernolle. Spectral estimation and normalisation for robust speech recognition. In *Proc. International Conference on Spoken Language Processing*, volume IV, pages 1997–2000, Philadelphia, U.S.A., October 1996.
[4] Kris Demuynck. *Extracting, Modelling and Combining Information in Speech Recognition*. PhD thesis, K.U.Leuven, ESAT, February 2001.
[5] Kris Demuynck, Jacques Duchateau, and Dirk Van Compernolle. Reduced semi-continuous models for large vocabulary continuous speech recognition in Dutch. In *Proc. International Conference on Spoken Language Processing*, volume IV, pages 2289–2292, Philadelphia, U.S.A., October 1996.
[6] Kris Demuynck, Jacques Duchateau, and Dirk Van Compernolle. A static lexicon network representation for cross-word context dependent phones. In *Proc. European Conference on Speech Communication and Technology*, volume I, pages 143–146, Rodos, Greece, September 1997.
[7] Kris Demuynck, Jacques Duchateau, and Dirk Van Compernolle. Optimal feature sub-space selection based on discriminant analysis. In Proc. European Conference on Speech Communication

and Technology, volume III, pages 1311–1314, Budapest, Hungary, September 1999.
[8] Kris Demuynck, Jacques Duchateau, Dirk Van Compernolle, and Patrick Wambacq. Improved feature decorrelation for HMM-based speech recognition. In Proc. International Conference on Spoken Language Processing, volume VII, pages 2907–2910, Sydney, Australia, December 1998.
[9] Kris Demuynck, Jacques Duchateau, Dirk Van Compernolle, and Patrick Wambacq. An efficient search space representation for large vocabulary continuous speech recognition. *Speech Communication*, 30(1):37–53, January 2000.
[10] Kris Demuynck and Tom Laureys. A comparison of different approaches to automatic speech segmentation. In Proc. 5th International Conference on Text, Speech and Dialogue, pages 277–284, Brno, Czech Republic, September 2002.
[11] Kris Demuynck, Tom Laureys, and Steven Gillis. Automatic generation of phonetic transcriptions for large speech corpora. In Proc. International Conference on Spoken Language Processing, volume I, pages 333–336, Denver, U.S.A., September 2002.
[12] Kris Demuynck, Tom Laureys, Dirk Van Compernolle, and Hugo Van hamme. Flavor: a flexible architecture for LVCSR. In *Proc. European Conference on Speech Communication and Technology*, pages 1973–1976, Geneva, Switzerland, September 2003.
[13] Kris Demuynck, Tom Laureys, Patrick Wambacq, and Dirk Van Compernolle. Automatic phonemic labeling and segmentation of spoken dutch. In *Proc. 4th International Conference on Language Resources and Evaluation*, volume I, pages 61–64, Lisbon, Portugal, May 2004.
[14] Kris Demuynck, Dirk Van Compernolle, and Patrick Wambacq. Doing away with the viterbi approximation. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 717–720, Orlando, U.S.A., May 2002.
[15] Ioannis Dologlou, Tom Claes, Louis ten Bosch, Dirk Van Compernolle, and Hugo Van hamme. Speaker normalization for automatic speech recognition – an on-line approach. In *Proc. EUSIPCO*, volume III, pages 1473–1476, September 1998.
[16] Jacques Duchateau. *HMM Based Acoustic Modelling in Large Vocabulary Speech Recognition*. PhD thesis, K.U.Leuven, ESAT, November 1998.
[17] Jacques Duchateau, Tim Ceyssens, and Hugo Van hamme. Use and evaluation of prosodic annotations in dutch. In *Proc. 4th International Conference on Language Resources and Evaluation*, volume IV, pages 1517–1520, Lisbon, Portugal, May 2004.
[18] Jacques Duchateau, Kris Demuynck, and Dirk Van Compernolle. Fast and accurate acoustic modelling with semi-continuous HMMs. *Speech Communication*, 24(1):5–17, April 1998.
[19] Jacques Duchateau, Kris Demuynck, Dirk Van Compernolle, and PatrickWambacq. Improved parameter tying for efficient acoustic model evaluation in large vocabulary continuous speech recognition. In *Proc. International Conference on Spoken Language Processing*, volume V, pages 2215–2218, Sydney, Australia, December 1998.
[20] Jacques Duchateau, Kris Demuynck, Dirk Van Compernolle, and Patrick Wambacq. Class definition in discriminant feature analysis. In *Proc. European Conference on Speech Communication and Technology*, volume III, pages 1621–1624, Aalborg, Denmark, September 2001.
[21] Jacques Duchateau, Kris Demuynck, and Patrick Wambacq. Discriminative resolution enhancement in acoustic modelling. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume III, pages 1245–1248, Istanbul, Turkey, June 2000.
[22] Jacques Duchateau, Kris Demuynck, and Patrick Wambacq. Confidence scoring based on backward language models. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 221–224, Orlando, U.S.A., May 2002.
[23] Jacques Duchateau and Patrick Wambacq. Unconstrained versus constrained acoustic normalisation in confidence scoring. In *Proc. International Conference on Spoken Language Processing*, volume III, pages 1617–1620, Denver, U.S.A., September 2002.
[24] Tom Laureys, Kris Demuynck, Jacques Duchateau, and PatrickWambacq. An improved algorithm for the automatic segmentation of speech corpora. In *Proc. 3rd International Conference on Language Resources and Evaluation*, volume V, pages 1564–1567, Las Palmas, Canary Islands, May 2002.
[25] Veronique Stouten, Hugo Van hamme, Jacques Duchateau, and Patrick Wambacq. Evaluation of model-based feature enhancement on the AURORA-4 task. In Proc. European Conference on Speech Communication and Technology, pages 349–352, Geneva, Switzerland, September 2003.
[26] Veronique Stouten, Hugo Van hamme, and Patrick Wambacq. Accounting for the uncertainty of speech estimates in the context of model-based feature enhancement. In Proc. International Conference on Spoken Language Processing, volume I, pages 105–108, Jeju Island, Korea, October 2004.
[27] Dirk Van Compernolle. Noise adaptation in a hidden Markov model speech recognition system. *Computer Speech and Language*, 3(2):151–168, 1989.
[28] Stefaan Van Gerven and Fei Xie. A comparative study of speech detection methods. In *Proc. European Conference on Speech Communication and Technology*, volume III, pages 1095–1098, Rodos, Greece, September 1997.
[29] Hugo Van hamme. PROSPECT features and their application to missing data techniques for

robust speech recognition. In Proc. International Conference on Spoken Language Processing, volume I, pages 101–104, Jeju Island, Korea, October 2004.
[30] Hugo Van hamme. Robust speech recognition using cepstral domain missing data techniques and noisy masks. In Proc. International Conference on Acoustics, Speech and Signal Processing, volume I, pages 213–216, Montreal, Canada, May 2004.
[31] Dong Hoon Van Uytsel. Probabilistic Language Modeling with Left Corner Parsing. PhD thesis, K.U.Leuven, ESAT, September 2003.
[32] Dong Hoon Van Uytsel, Filip Van Aelten, and Dirk Van Compernolle. A structured language model based on context-sensitive probabilistic left-corner parsing. In Proc. 2nd Meeting of the North American Chapter of the Association for Computational Linguistics, pages 223–230, Pittsburgh, PA, USA, June 2001.
[33] Dong Hoon Van Uytsel and Dirk Van Compernolle. Language modeling with context-sensitive probabilistic left corner parsing. Computer Speech and Language, 19(2):171–204, April 2005. url: http://authors.elsevier.com/sd/article/S0885230804000221.
[34] Dong Hoon Van Uytsel, Dirk Van Compernolle, and PatrickWambacq. Maximum-likelihood training of the PLCG-based language model. In Proc. IEEE Automatic Speech Recognition and Understanding Workshop 2001, Madonna di Campiglio, Italy, December 2001. 4 pages. ISBN 0-7803-7343-X.
[35] Fei Xie and Dirk Van Compernolle. Speech enhancement by spectral magnitude estimation — a unifying approach. *Speech Communication*, 19(2):89–104, August 1996.