The Joint LIMSI and Vecsys Research Systems for NBEST 2008

Julien Despres[2], *Petr Fousek*[1], *Jean-Luc Gauvain*[1], *Sandrine Gay*[2],
Yvan Josse[2], *Lori Lamel*[1], *Abdel Messaoudi*[1,2]

[1]Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, France
{gauvain,lamel,fousek,abdel}@limsi.fr

[2]Vecsys Research
3, rue Jean Rostand
91400 Orsay, France
{despres,gay,josse}@vecsysresearch.com

Abstract

This document describes the speech recognizers jointly submitted by the LIMSI and Vecsys Research to the N-Best 2008 evaluation. The aim of this evaluation was to perform automatic speech recognition (ASR) for the Dutch language. Northern Dutch and Southern Dutch (also known as Dutch - NL, and Flemish - VL) have been processed with two different data types per accent (broadcast news - BN, and conversational telephone speech - CTS). The speech recognizers use multiple decoding passes with models (lexicon, acoustic models, language models) trained for the four different transcription tasks: BN-NL, BN-VL, CTS-NL and CTS-VL. Four primary systems (one for each accent-domain task) have been trained for the primary training condition and the unlimited decoding. The primary submission is also a less than 10xRT contrastive submission. Four contrastive systems have a processing time of 1xRT. The case-sensitive word error rates (WER) of the primary LIMSI-Vecsys Research systems on the N-Best development data are 9.5% for BN-NL, 8.7% for BN-VL, 31.6% for CTS-NL and 41.9% for CTS-VL.

Index Terms: automatic speech recognition, Dutch, Flemish, CGN, broadcast news, conversational telephone speech, MLP, PLP, MMIE, SAT, ROVER.

1. Introduction

N-Best 2008 aims at setting up the infrastructure for a benchmark evaluation in large vocabulary speech recognition for the Dutch language. The evaluation is conducted by TNO Human Factors Soesterberg, the Netherlands in co-operation with Spex in Nijmegen. The evaluation framework can serve both as a basis for future evaluations, which can probe the progress in large vocabulary speech recognition for Dutch, and serve as an aid for the development of new speech recognition technologies for the Dutch language.

Two large vocabulary speech recognition tasks are covered in the N-Best 2008 evaluation data: Broadcast News (BN) and Conversational Telephone Speech (CTS). Two main dialect regions have also been defined: Northern and Southern Dutch, as spoken by people from The Netherlands and from Flanders (Belgium), respectively. The participants to the benchmark should use a common speech database, the Corpus Gesproken Nederlands (CGN) for acoustic training of their primary systems, as well as other common resources for language modeling and pronunciation modeling.

Speech	*Duration (hours)*		*Total words*	
	NL	VL	NL	VL
BN	99.4 / 84.0	52.9 / 48.0	1.1M	572.2K
CTS	92.0 / 80.0	64.0 / 60.0	1.3M	808.3K

Table 1: *N-Best acoustic data (total / primary training) provided by CGN. The total data contains both the primary training data and the development data.*

Language	*Epoch*	*Corpus (M words)*	*Voc. (M words)*
NL	1999-2004	360.0	7.2
VL	1999-2004	1418.2	14.8

Table 2: *Primary N-Best language training data (provided by PCM and Mediargus).*

2. Task and data description

Baseline acoustical and language modeling training data was provided by TNO. The acoustic training material shown in Table 1 was entirely obtained from CGN (Corpus Gesproken Nederlands, Spoken Dutch Corpus). The audio files and the corresponding transcriptions were separated into Dutch and Flemish parts, including BN and CTS components for each dialect. About 100 hours ($\sim$1.2M words) and 55 hours ($\sim$700K words) of audio recordings were available for BN-NL and BN-CTS on one hand, and BN-VL and CTS-VL on the other hand. Since the development data sets were included in the CGN data, they had to be removed from training. The language modeling training data presented in Table 2 (broadcast news only) was composed of newspaper articles from 1999 to 2004, obtained from the Dutch publisher PCM and the Flemish Mediargus. The data contain approximately 360M words for the Dutch portion and 1418M words for the Flemish subset.

Although further corpora could be added provided that their creation date predates 1 January 2007, no additional data was used for the the LIMSI-Vecsys Research speech recognizer training.

Table 3 summarizes the development and the evaluation data. Each time all the segments were entirely decoded then a UEM file was used to select the segments that would finally be scored. For the CTS audio files the conversations were recorded on two different channels, each channel was consequently separately decoded (an overlap phenomenon sometimes appeared in the development data, disturbing the decoding). The develop-

Task	*Data Type*	*Total Duration (h)*	*Scored Duration (h)*	*Scored #words*
BN-NL	Dev.	1.1	1.0	8721
	Eval.	9.0	2.2	-
BN-VL	Dev.	1.0	1.0	10406
	Eval.	5.4	2.1	-
CTS-NL	Dev.	2.0 (x2 ch.)	1.8	6695
	Eval.	5.7 (x2 ch.)	2.9	-
CTS-VL	Dev.	1.9 (x2 ch.)	1.8	6790
	Eval.	6.5 (x2 ch.)	2.5	-

Table 3: *Development and evaluation data. The total duration corresponds to the length of the audio files. The scored duration corresponds to the duration of the segments given by the UEM file (only these segments, which size in words is given, is scored).*

ment files were composed of CGN excerpts, except for the BN-NL task which also included parts from an other data source. In summary, for the development data, 1 hour ($\sim$9K words) was scored for each BN task and a bit under 2 hours ($\sim$7K words) for each CTS task. The evaluation data set is larger, with between 2 and 3 hours of data to be scored for each of the 4 tasks.

3. Speech recognizer overview

The BN speech recognizers for Dutch and Flemish use the same basic modeling and decoding strategy as in the LIMSI English broadcast news system [8]. The acoustic and language models are language and task specific. As for the dictionaries, the Dutch and Flemish ones use the same word list and pronunciation variants but the pronunciation probabilities collected during the acoustic training are task-specific. All the LIMSI-Vecsys Research systems are only trained on the TNO data. Table 4 lists all the submitted systems.

The primary recognition submission results from a ROVER [4] between two system outputs, using different acoustic features: PLP and MLP, each one being generate in 2 decoding passes. Each of these systems include rescoring by a 4-gram neural network LM. The PLP systems for both BN and CTS reuse the 1xRT output as a first pass to adapt the acoustic models. Unsupervised acoustic model adaptation is also used in the CTS MLP system (also with the 1xRT system hypotheses), but adaptation is not performed in the BN MLP system.

The LIMSI-Vecsys Research primary speech recognizers process the audio data in less than ten times RT, meaning that the primary systems are the same as the contrastive systems for the "less than 10xRT" condition.

The 1xRT word recognition is performed in a single decoding pass, using a 2-gram LM for decoding an 4-gram LM for rescoring.

4. Audio partitioner

The broadcast news audio partitioner is based on an audio stream mixture model [7, 8]. First, the non-speech segments are detected and rejected using Gaussian mixture models (GMMs) representing speech, speech over music, noisy speech, pure-music and other background conditions. An iterative maximum likelihood segmentation/clustering procedure is then applied to the speech segments. The result of the procedure is a sequence of non-overlapping segments with their associated segment cluster labels. The objective function is the GMM log-likelihood penalized by the number of segments and the number of clusters, appropriately weighted. Four sets of GMMs are then used to identify telephone segments and the speaker gender.

The CTS audio partitioner uses the same basic method to divide the acoustic signal into homogenous segments, however the segmentation/clustering step is not needed since all speech segments are assumed to come from the same speaker. Non-speech segments are detected and rejected using telephone-band GMMs representing speech and silence, where the silence model represents the background conditions.

Task	*Processing time*	
	Primary 10xRT	*Contrast 1xRT*
BN	2-pass PLP $\oplus$ 1-pass MLP	1-pass PLP
CTS	2-pass PLP $\oplus$ 2-pass MLP	1-pass PLP

Table 4: *Summary of the speech recognizer characteristics for the Primary and Contrast submissions ($\oplus$ means "ROVER"). The contrastive 1xRT PLP system output is also used as the first pass of the primary 10xRT PLP system.*

5. Acoustic features and models

Two sets of features are used for each task. The first are standard cepstral features (perceptual linear prediction - PLP), and the second, cepstral features produced with a multi layer perceptron (MLP) [19, 6]. The MLP features are based on a recently proposed Bottle-Neck architecture [11] with long-term warped LP-TRAP speech representation at the input.

The PLP feature vector has 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band every 10ms (0-3.8kHz band for CTS). For each 30ms frame the Mel scale power spectrum is computed, and the cubic root taken followed by an inverse Fourier transform. Then LPC-based cepstrum coefficients are computed. The cepstral coefficients are normalized on a segment-cluster basis using cepstral mean removal and variance normalization. Thus each cepstral coefficient for each cluster has a zero mean and unity variance. The 39-component acoustic feature vector consists of 12 cepstrum coefficients and the log energy, along with the first and second order derivatives.

The MLP features are generated in two steps. First raw features, typically with a wide temporal context of 100–500 ms, are extracted and input to the MLP. These features are then processed by the MLP followed by a principal component analysis (PCA) transform to yield the hidden Markov models (HMM) features. Time-warped linear predictive TRAP (wLP-TRAP) [5] features are used. Separate MLPs were trained for each task and dialect, using 180 state targets (one for each state of the 38 phones, and one state for each non-phone unit) using the training scheme described in [6]. The MLP features are then concatenated with the PLP features resulting in a 78 component feature vector.

All acoustic models (AMs) are tied-state, left-to-right context-dependent, HMMs with Gaussian mixtures. The triphone-based context-dependent phone models are word-independent but position-dependent. The tied states are obtained by means of a decision tree. Different sets of gender-

independent acoustic models were trained for each task (BN and CTS), and each dialect (Northern and Southern Dutch).

The models all use speaker-adaptive (SAT) and Maximum Mutual Information Estimation (MMIE) training. For each task and dialect, models were trained using both standard PLP and concatenated MLP+PLP features. For the PLP models, a maximum-likelihood linear transform (MLLT) is also used, but not for the concatenated MLP+PLP models.

The BN and CTS model sets cover about 22k and 20k phone contexts, respectively, with 11.5k tied states and 32 Gaussians per state. Silence is modeled by a single state with 1024 Gaussians. Initially dialect independent models are trained on all of the available data for the task, that is about 130 hours for BN and 150 hours for CTS. These models serve as priors for Maximum *a Posteriori* (MAP) [10] estimation of dialect-specific models.

6. Pronunciation lexicons

Pronunciations are based on a 41 phone set (16 vowels, 22 consonants and 3 other symbols that represent silence, filler words, and breath noises). These phones, listed in Table 5, are the most common in the Dutch/Flemish language. Short and long vowels are differentiated, common diphthongs are written with one phone symbol (as opposed to a sequence of phones), as well as the hard and soft pronunciations of the Dutch "g/ch" graphemes. Infrequent phones used in loan words (for example, nasalized vowels) were not included in the phone set. The pronunciations are encoded using one character code per phone to simplify readability by humans and the inclusion of multiple pronunciations. The table gives the VR phone symbol along with the Sampa code, the Dutch graphemes and an example word.

Two master dictionaries served as basis to generate the lexicons used in the transcription tasks. The first one is the Dutch master dictionary, based on the CELEX [1] dictionary and the Dutch part of the CGN dictionary. The second one, the Flemish master dictionary, is derived from the Flemish part of the CGN dictionary and the FONILEX [15] dictionary.

A lexicon is formed for a list of words (see Section 7) and by extracting the pronunciations for these words from the master lexicon (Dutch or Flemish). Words for which no pronunciation is present in the master lexicon are phoneticized by a statistical approach using the translation tools Giza++ [16] and Moses [12]. This approach was inspired by the method described by Walter Dealemans and Antal van den Bosch [3]. With this method, multiple pronunciations are generated for a given word, and the best *n* in terms of probability are kept. Particular pronunciations are also added for some classes of words (acronyms and proper nouns). An acronym can be pronounced as a word or can be spelled. An additional English pronunciation is given for most proper nouns. Initially two lexicons were generated – one Dutch-oriented, the other Flemish-oriented – and then merged into one.

The characteristics of the recognition lexicons are summarized in Table 6. Two large dictionaries containing 300k and 500k entries cover the two languages involved in this evaluation. Task-oriented versions of the dictionaries were created by enriching the merged ones with pronunciation counts (BN/CTS).

7. Language modeling

Word lists for Dutch and Flemish were selected by choosing all the words in the transcriptions of the training portion of the au-

Phone		*Dutch*	
VR	*Sampa*	*Grapheme*	*Example*
		Vowels	
I	I	i	bit
i	i:	ie,i	biet
Ü	Y	u	hut
ü	y:	uu,u	fuut
E	E	e	bed
e	e:	ee,e	beet
ø	@	e	de
ö	|:	eu	neus
A	A	a	bad
a	a:	aa,a	baad
O	O	o	bot
o	o:	oo,o	boot
u	u:	oe	hoed
é	EI	ij,ei	bijt, ei
ó	\I	ui	buit
à	Au	ou,ouw	bout
		au,auw	
		Consonants	
p	p	p	pen
b	b	b	biet
t	t	t,d	tak
d	d	d	dak
k	k	k,c	kat
g	g	g	goal
m	m	m	mens
n	n	n	nek
Ñ	N	ng	eng
f	f	f	fiets
v	v	v,w	oven
s	s	s,c	sok
z	z	z	zeep
S	S	ch,sj	chef
Z	Z	j	jury
X	x	ch,g	acht
G	g	g	gaan
r	r	r	rat
h	h	h	hoed
w	w	w	wang
j	j	j	jas
l	l	l	land
		Others	
.		silence	
®		breath	
&		filler word	

Table 5: *The Dutch phone set (38+3 symbols).*

#words	*300K*	*500K*
Language	NL+VL	NL+VL
#phones	41	41
#nonspeech	3	3
prons per word	4.37	4.91

Table 6: *Recognition lexicons. For each word list, separate lexicons are generated for each dialect, and the two are merged.*

Task	*#words*	*OOV*	*#(2,3,4)g*	*4g ppx*
BN-NL	300K	0.8%	(45M, 15M, 4.9M)	254.0
	500K	0.6%	(49M, 15M, 4.9M)	253.1
BN-VL	300K	0.7%	(54M, 21M, 8.3M)	213.9
	500K	0.6%	(58M, 22M, 8.2M)	213.6
CTS-NL	300K	0.5%	(18M, 5.2M, 1.5M)	91.7
CTS-VL	300K	0.5%	(15M, 4M, 1M)	112.5

Table 7: *Summary of language model development. All models were generated using a cut-off of 1-2-3 and a pruning value of 1e-10.*

dio data and the most frequent words in the text corpus regardless of the dialect. Therefore the word lists are the same for both dialects. The size of the vocabulary (n) was chosen to minimize the OOV rate on the four development data sets (while keeping a reasonable number of words with regards to the decoding speed). A first 300K case-sensitive word list was chosen, yielding an OOV rate under 1% for all development data. A second 500K case-sensitive word list was generated, with an OOV rate close to 0.5% on all development sets.

The texts were normalized to a common form. To facilitate the text normalization the transcriptions and the newspaper articles were processed separately. No special treatment was applied to convert the written texts closer to a spoken form, and all language models were estimated on the same normalized text corpus for the four tasks.

Text normalization entails multiple steps. First, identical articles were removed. Then numerical expressions were treated ("497,2 miljoen euro" becomes "vierhonderdzevenennegentig komma twee miljoen euro"). Since the capitalization of words is scored, a step was added to properly re-case all of the texts. The pseudo-compounded words (i.e., words with a dash) were separated but the dash was kept in the text, either alone or joined to the previous or following word. The apostrophes were kept agglutinated to the words except in some cases ("d'rachter" becomes "d'r achter", "euro's" becomes "euro 's"). The texts were finally split into sentences and the main punctuation was removed. After processing, the number of words available was about 3.7M words in the transcriptions and 1.5G words in the text articles, with a global vocabulary size of about 6M words. In order to build the language models the transcriptions were split into subsets by task and language: i.e., separate parts for BN-NL, BN-VL, CTS-NL, and the CTS-VL transcriptions. The articles were also split according to source (ie: Algemeen Dagblad, De Morgen, De Standard, etc.).

For all systems, n-gram language models were obtained by interpolation of backoff n-gram language models using the modified Kneser-Ney smoothing (as implemented in the SRI toolkit [18]) trained on separate subsets of the available language model training texts. The characteristics of the language models are summarized in Table 7. The language models result from the interpolation of component LMs trained on 26 sources:
1) Audio transcriptions (4 sources, one for each task): 3.8M words (cut-off 0-0-0).
2) NL web texts (10 sources): 357M words (cut-off 0-1-2)
3) VL web texts (12 sources): 1215M words (cut-off 0-1-2)
The mixture weights were automatically chosen by an EM algorithm to minimize the perplexity of the development data. The 2-gram models models used for decoding were heavily pruned and contain fewer than 1M 2-grams. The 4-gram models were pruned with a coefficient of 1e-10 and contain about 5M 4-grams for BN-NL, 8M for BN-VL, 1M for CTS-NL and 1.5 for CTS-VL. The perplexity obtained on the BN-NL, BN-VL, CTS-NL and CTS-VL development data sets are respectively 254.0, 253.1, 91.7 and 112.5.

Task	*System*	*Decoding pass*	
		Pass1	*Pass2*
BN-NL	PLP	11.9	10.0
	MLP	10.3	-
BN-VL	PLP	11.6	9.1
	MLP	9.3	-
CTS-NL	PLP	37.8	33.2
	MLP	36.8	33.7
CTS-VL	PLP	48.8	45.5
	MLP	46.0	42.5

Table 8: *Case sensitive WER (in %) after each decoding pass on the dev08 development data for PLP and MLP systems. Punctuation and non-lexical events are not scored.*

8. Decoding

Word recognition is performed with two distinct systems, each using one or two decoding passes. The first system uses a classical PLP signal analysis whereas the second uses a MLP analysis. Each decoding pass of the two systems 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 [13]. The lattices produced in the last pass of both systems are rescored by the neural network LM interpolated with a 4-gram back-off LM. Then, a ROVER combination of the two systems is carried out. More specifically, the decoding steps are:
1) Initial hypothesis generation using large MLLT and MMIE-trained AMs ($\sim$ 1.0xRT). The submission for the 1xRT condition is the result of this first pass.
2) Multiple-class MLLR adaptation of first pass AMs, followed by a rescoring of the produced lattices with a neural network interpolated with 4-gram back-off LM. A decision tree is used to determine the number of MLLR transforms given the available adaptation data and the tied states associated to each regression class. Tables 8 and 9 give the word error rates on the NBEST dev08 data. For the primary system, which also serves as an under 10xRT submission, the word error rates are 9.5% for BN-NL, 8.7% for BN-VL, 31.6% for CTS-NL and 41.9% for CTS-VL. If case is not scored (that is if case differences are not counted as errors), the WER decreases by about 1% on average for the BN tasks. As can be seen in the table, ignoring case only reduces the WER on the CTS tasks by about 0.2%.

Table 10 gives the word error rates for a slower CTS system that runs in under 20xRT. An intermediary adaptation pass using a single MLLR class has been inserted between the first and second pass of the PLP based system. The 3-pass PLP based system achieves a WER reduction of 0.1% for NL and 0.9% for VL (compare the pass 3 results in Table 10 to the pass 2 results in Table 8). For the MLP based system, a slower second pass decoding is carried out, which results in a WER reduction

Task	*System*	
	Primary (10xRT)	*Contrastive (1xRT)*
BN-NL	9.5 / 8.2	11.9 / 10.6
BN-VL	8.7 / 7.8	11.6 / 10.7
CTS-NL	31.6 / 31.4	37.8 / 37.6
CTS-VL	41.9 / 41.7	48.8 / 48.6

Table 9: *Final case-sensitive/case-insensitive WER (in %) on the dev08 development data for the 4 tasks for the primary (and also less than 10xRT:* ROVER *between the last passes of PLP and MLP) and the contrastive systems (1xRT: PLP-pass1). Punctuation and non-lexical events are not scored. If case differences are not counted as errors, the WER decreases by 1% on average for the BN tasks.*

Task	*System*	*Decoding pass*			ROVER
		Pass1	*Pass2*	*Pass3*	
CTS-NL	PLP	37.8	35.4	33.1	31.1
	MLP	36.8	32.5	-	
CTS-VL	PLP	48.8	45.8	44.6	41.0
	MLP	46.0	41.6	-	

Table 10: *Case sensitive WER (in %) for the CTS data after each decoding pass on the dev08 development data for PLP and MLP systems. Punctuation and non-lexical events are not scored.*

of 1.2% and 0.9% for NL and VL respectively. The rightmost columns gives the ROVER result for the 3-pass PLP system and the 2-pass MLP system. Compared to the CTS results in Table 9 the word error rate for NL is reduced by 0.5% (from 31.6% to 31.1%) and by 0.9% (from 41.9% to 41.0%) for VL. Scoring without case distinction reduces the word error rates to 31.0% and 40.8% respectively.

9. Summary

This document has given an overview of the speech recognizer and the task-specific models used in the joint submissions by LIMSI and Vecsys Research for the Dutch N-Best 2008 evaluation 9. The submissions for the baseline condition are the same as the under 10 times real-time contrast system. A second set of contrastive results were submitted for a real-time system. In total 8 systems were developed (1xRT and 10xRT), for each dialect (Northern and Southern) and task (BN and CTS). Dialect-specific acoustic models for each task were obtained by MAP adaptation of dialect-independent models trained on all the available data for the task. Different language model interpolation coefficients were used for the different conditions. Word error rates under 10% were obtained on broadcast news development data, and on the order of 30% for Dutch and 40% Flemish conversational data. It is difficult to know if these results will extrapolate to the evaluation data, which, according to the organizers is not well-represented by the development data.

10. References

[1] Baayen, R.H., Piepenbrock, R., and Gulikers, L., "The CELEX lexical database" (CDROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania, 2005.

[2] Barras, C., Zhu, X., Meignier, S., Gauvain, J.L., "Multi-stage speaker diarization of broadcast news," *IEEE Trans. on Audio, Speech and Language Processing*, 2006.

[3] Daelemans, W. and van den Bosch, A., "A language-independent, data-oriented architecture for grapheme-to-phoneme conversion", *Proceedings of the ESCA-IEEE conference on Speech Synthesis*, New York, 1994.

[4] Fiscus, J.G., "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," *Proceedings ASRU*, 1997

[5] Fousek, P., *Extraction of Features for Automatic Recognition of Speech Based on Spectral Dynamics*, Prague: PhD thesis, Czech Technical University in Prague, Faculty of Electrical Engineering, 2007.

[6] Fousek, P., Lamel, L. and Gauvain, J.L., "Transcribing Broadcast Data Using MLP Features," submitted to ICSLP'08.

[7] Gauvain, J.L., Lamel, L., Adda, G., "Partitioning and Transcription of Broadcast News Data," *ICSLP'98*, 5:1335-1338, 1998.

[8] Gauvain, J.L., Lamel, L. and Adda, G., "The LIMSI Broadcast News Transcription System," *Speech Communication,* 37(1-2):89-108, 2002.

[9] Gauvain, J.L., Lamel, L., Schwenk, H., Adda, G., Chen, L. and Lefevre, F., "Conversational telephone speech recognition," *IEEE ICASSP'03*, I:212-215, Hong Kong, 2003.

[10] Gauvain, J.L., Lee, C.H., "Maximum *a Posteriori* Estimation for Multivariate Gaussain Mixture Observation of Markov Chains," *IEEE Trans. on Speech and Audio Processing*, **2**(2):291-298, 1994.

[11] Grézl, F. and Fousek, P., "Optimizing bottle-neck features for LVCSR," ICASSP'08, Las Vegas, ND, 2008.

[12] Koehn, P. and Hoang, H., et al., "Moses: Open Source Toolkit for Statistical Machine Translation," 2007.

[13] Leggetter, C.J., Woodland, P.C., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, **9**(2):171-185, 1995.

[14] Mangu, L., Brill, E., Stolcke, A., "Finding Consensus Among Words: Lattice-Based Word Error Minimization," *ISCA EuroSpeech'99*, 495-498, Budapest, 1999.

[15] Mertens, P., Vercammen, F., "The Fonilex Manual," Centre for Computational Linguistics, K.U.Leuven, 1997.

[16] Och, F.J. and Ney, H., "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, **29**(1):19-51, 2003.

[17] Schwenk, H., "Continuous space language models", *Computer Speech & Language*, **21**:492-518, 2007.

[18] Stolcke, A., "SRILM – an extensible language modeling toolkit," 2002.

[19] Zhu, Q., Stolcke, A., Chen, B.Y. and Morgan, N., "Using MLP features in SRI's conversational speech recognition system," *InterSpeech'05*, pp. 2141-2144, 2005.