

The ESAT 2008 System for N-Best Dutch Speech Recognition Benchmark

Kris Demuynck ^{#1}, Antti Puurula ^{#2}, Dirk Van Compernelle ^{#3}, Patrick Wambacq ^{#4}

[#] *Department of Electrical Engineering, Katholieke Universiteit Leuven
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium*

¹ *kris.demuynck@esat.kuleuven.be*

² *antti.puurula@esat.kuleuven.be*

³ *dirk.vancompernelle@esat.kuleuven.be*

⁴ *patrick.wambacq@esat.kuleuven.be*

Abstract—This paper describes the ESAT 2008 Broadcast News transcription system for the N-Best 2008 benchmark, developed in part for testing the recent SPRAAK Speech Recognition Toolkit. ESAT system was developed for the Southern Dutch Broadcast News subtask of N-Best using standard methods of modern speech recognition. A combination of improvements were made in commonly overlooked areas such as text normalization, pronunciation modeling, lexicon selection and morphological modeling, virtually solving the out-of-vocabulary (OOV) problem for Dutch by reducing OOV-rate to 0.06% on the N-Best development data and 0.23% on the evaluation data. Recognition experiments were run with several configurations comparing one-pass vs. two-pass decoding, high-order vs. low-order n-gram models, lexicon sizes and different types of morphological modeling. The system achieved 7.23% word error rate (WER) on the broadcast news development data and 20.3% on the much more difficult evaluation data of N-Best.

I. INTRODUCTION

The N-Best project [1] was one of the projects within the STEVIN program [2] and had the goal to establish a benchmark for Dutch Large Vocabulary Continuous Speech Recognition (LVCSR) and to evaluate the performance of present-day LVCSR systems for the Dutch language. The N-Best evaluation consisted of four subtasks: transcription of Broadcast News (BN) and Conversational Telephone Speech (CTS) for both Northern (the Netherlands) and Southern Dutch (Belgium).

ESAT participation in N-Best Evaluation consisted of providing the SPRAAK [3] recognition toolkit to other participants, and proceeding development of the system for the Southern Dutch BN sub-task presented here. Subsequently acoustic modeling and the decoder represent largely the standard methods in current LVCSR research as supported by SPRAAK. The areas improved were text normalization, pronunciation modeling, lexicon selection and morphological modeling with the intention of correcting the Out-of-Vocabulary (OOV) problem for Dutch. In addition, Real Time Factor (xRT) and Word Error Rate (WER) were investigated with several configurations with one-pass vs. two-pass decoding, lexicon size, n-gram order and use of morphological modeling.

In general successful BN recognition requires a system capable of acoustic segmentation, speaker clustering and speaker adaptation, noise robustness and large vocabularies. While BN is a challenging area of LVCSR research on its own, Dutch LVCSR has some complicating factors due to strong dialects resulting in pronunciation and vocabulary variation, and morphological productivity resulting in a very large number of compound words.

II. TASK DESCRIPTION

A. N-Best 2008 Dutch Datasets

Datasets for acoustic training in N-Best evaluation consisted of Southern and Northern Dutch BN and Continuous Telephone Speech (CTS) components from the Corpus Gesproken Nederlands (CGN) [4]. Text data consisted of the corresponding audio transcriptions along with Dutch newspaper texts. In summary the training and development datasets consisted of:

- Audio data, after filtering:
 - 39.9/63.0 hours of Southern Dutch BN/CTS audio
 - 75.6/90.9 hours of Northern Dutch BN/CTS audio
 - 1 to 2 hours of development data for each subtask
- Text data, after normalization and filtering:
 - 1104m words from 12 Southern Dutch newspapers
 - 356m words from 10 Northern Dutch newspapers
 - 3.7m words from the CGN audio transcriptions

The evaluation data consisted of 2 to 3 hours of audio for each subtask. In contrast to the N-Best evaluation guidelines the evaluation data was very different from the training and development data. The Southern Dutch BN evaluation data neither had a significant portion of telephone speech segments in contrast to the development data. For most N-Best participants the differences meant WERs that were roughly doubled on the evaluation data, as audio normalization and adaptation were not the foci of development.

N-Best scoring requires standardized output including number formatting, correct compounding and capitalization of proper nouns and acronyms. Punctuation, filled pauses, hesitations and other non-lexical sounds are ignored if the recognizer identified them as such. Remapping rules are used to allow

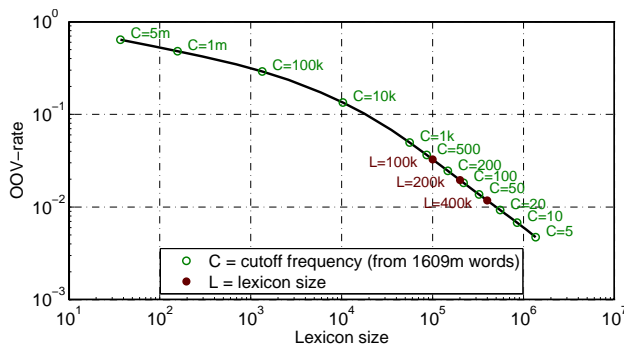


Fig. 1. OOV-rate on training data for different cutoffs and lexicon sizes

equivalent alternatives such as transcriptions of foreign names. A number of errors in the reference transcriptions of both the evaluation and development set were corrected, including capitalization, spelling errors and wrong transcriptions.

B. LVCSR for Dutch

Dutch is a morphologically productive language that is rich in the use of compounding to produce new words. Situated between the major European languages, Dutch also has a large number of loan words and an ongoing spelling reform resulting in a considerable Out-of-Vocabulary (OOV) rate due to untypical causes. This is illustrated in Figure 1, where OOV-rates are shown for after text normalization for lexica of different frequency cutoffs and the word lexica used in this paper.

TABLE I
CLASSIFICATION OF OOV WORDS FOR THREE LEXICON SIZES

OOV type	Lexicon size		
	100k	200k	400k
Proper noun	52.7%	52.8%	51.5%
Compound	34.5%	37.9%	41.8%
Inflection/derivation	5.6%	3.5%	2.9%
New word	3.7%	2.1%	1.1%
Abbreviation	2.3%	2.3%	1.3%
Foreign word	1.2%	1.4%	1.3%

As each OOV word results in around 2 to 3 errors in recognition, it is important to keep the OOV-rate in a recognizer to a minimum. Table I gives a further classification of the OOV words excluding spelling errors on a random sample of 3500 OOV words for a 100k lexicon, of which 2399 were OOV for 200k and 1721 for 400k. As can be seen, the OOV-rate in Dutch is due to different causes. Ways to manage the OOV-rate are given in the following sections.

Dutch LVCSR is also complicated by strong dialectal variation. The most obvious difference between Northern and Southern Dutch is a significant difference at the pronunciation level. This issue was addressed by using phonological rules optimized on development data, yielding a median number of pronunciations per word of 3.8. In addition for acoustic modeling Northern and Southern Dutch varieties were considered separate, and only the Southern Dutch data was used. For the

written language the differences were considered minimal, and data from both varieties was used.

III. ACOUSTIC MODELING AND DECODER

A. Acoustic Modeling

SPRAAK [3] was used for training acoustic models (AM) on the Southern Dutch audio data. The AMs were triphone Hidden Markov Models (HMM) with state emissions modeled by Gaussian Mixture Models (GMM) with globally tied Gaussians. Two AMs were trained: one on the BN data to handle wideband speech and one on CTS data to handle narrowband telephone speech that occurs during news shows. The BN data was filtered to remove telephone interviews and noisy segments. In the CTS data segments with crosstalk were removed.

MIDA features [5] were used for acoustic modeling. These were computed by applying a mutual information based discriminant linear transform on mean-normalized and vocal-tract length normalized [6] (VTLN) mel-scale spectral features with first and second order time derivatives. This reduced 66 dimensional feature vectors to 25 dimensions for CTS and 36 for BN. Initial segmentation was done with pre-existing models using the Dutch YAPA phone set [7]. Acoustic units consisted of 49 three-state cross-word triphones (46 phones, silence, garbage and speaker noise) and one single-state triphone (short schwa). The cross-word triphone states were tied with a decision tree tying algorithm [8], and the density for each of the 4k states is modeled with a GMM with a global pool of 50k (BN) or 65k (CTS) Gaussians.

B. Speaker Segmentation and Clustering

Speaker clustering and segmentation was done prior to spectral mean-normalization and VTLN. An initial segmentation of the audio into sentence-like chunks was done with a 6-state HMM comprising a silence, male and female state for wideband and telephone speech. Speaker clustering was done by spectral clustering [9] on a similarity matrix comparing all speech segments. The similarity matrix incorporated information such as spectral similarities, temporal proximity, and initial classification. Speaker segmentation and clustering ran in less than $0.025 \times \text{RT}$ and provided perfect wideband/telephone classification and adequate speaker clustering on the development data. As only mean normalization and VTLN were used, using speaker clustering decreased the WER on the development data by a mere 0.2% absolute over non-clustered adaptation.

C. Decoding

SPRAAK uses a time-synchronous decoder designed to integrate complex knowledge sources including high-order n-grams and cross-word context-dependent phones in a single pass. This is done by precompiling a pronunciation dictionary to a state-emission Finite State Transducer (FST) by tying triphone states and organizing the language model (LM) into a prefix tree. Viterbi decoding is implemented by integrating LM probabilities dynamically with a token-passing algorithm [10],

[5] into the static search network. Decoding is accelerated by smearing unigram probabilities, caching LM lookups, adaptive beam pruning [5] and Gaussian selection [11].

Having a separation between the major components (AM, pronunciation dictionary and LM) allows the use of optimized precompiled representations for each component. The largest configuration used 41MB for the AM, 64MB for the pronunciation dictionary of 400k pronunciation lattices and 2.8GB for the 6-gram LM of 395m n-grams.

The decoder hypothesized sentence boundaries on pauses. On the development data set this decreased the WER by 1.2% absolute. In addition non-speech noise and speech noise states were used in recognition. Two alternative methods for modeling cross-word coarticulation were tested on the development data. The first one was based on the assumption that cross-word coarticulation is phonemic in nature, and used cross-word rewrite rules as in [7] to model word-boundary effects. The second assumed that phones should take word position into account, and used position dependent phones [12]. These methods didn't give improvement on the development data and were discarded from further experiments.

IV. LEXICON AND LANGUAGE MODELING

A. Text Normalization and Filtering

Effort was put into normalizing text data into a form that is most suitable for language modeling, and it is our belief that considerable improvements can be gained from detailed text normalization. Definition of a normalized form gives LM training matching data and simplifies many subsequent steps such as pronunciation dictionary generation. Our data normalization involved several separate steps: (proper) text normalization, spelling correction and data filtering. This separation and order of steps is important as the subsequent steps only need to correct certain types of abnormalities in the text.

Text normalization was done with cascaded regular expression replacements to get a normalized form by conversion of characters, sentence boundary marking, formatting acronyms and expanding web addresses, numbers, 965 abbreviations and 107 measure words. Incorrect short words were filtered with a stoplist, as these make decoding slower. Trash removal was done between the above steps by filters that were designed on the basis of text statistics before and after the steps.

Spelling correction handled capitalization, hyphenation errors and correction of 4340 spelling variants. First two of these involved unsupervised use of within-corpus statistics to correct words incorrectly capitalized or split with hyphens. The spelling correction was improved by unsupervised detection and manual verification of spelling variants based on similarity of bigram statistics of words with Levenshtein distance two or less, and rule-based morphological expansion of corrections for the found variants.

Finally the data was filtered by removing duplicate sentences and sentences with a high rate of uncommon words. Filtering was done to each of the 26 components for lexicon selection and to 4 main text components for language mod-

eling: the 12 Southern Dutch newspapers, 10 Northern Dutch newspapers and the BN and CTS transcripts.

B. Pronunciation Dictionary Generation

Pronunciation dictionary generation was handled by an updated version of the system described in [13]. The core lexicon is Fonilex [7] which provides multiple Southern Dutch phonemic transcriptions for 170k common Dutch words. Words not found in a precompiled lexicon are piped to the following modules in succession:

- A rule-based inflection, derivation and compounding module that finds decompositions and merges pronunciations using assimilation rules.
- A module that handles acronyms.
- A grapheme to phoneme (g2p) module using the ID3 algorithm [14], trained on the Fonilex lexicon.

Fonilex provides a rule set for generating pronunciation variants. This set was extended to generate all likely variants and optimized on WER on the development data. This resulted in a median of 3.8 pronunciations per word or 1.13 variants per phone. One of the deficiencies of this system, as in many other LVCSR systems, is that pronunciation generation for proper nouns remained inaccurate, particularly for foreign names.

C. Morphological Modeling

To reduce the OOV-rate caused by mostly compounding, three alternatives to word-based language models for Dutch were compared. As a basic solution fixing compounds with post-processing was tested, this involved using a large lookup table for compounding recognition results. Other solutions were one based on decompounding words for LM training, and one on using language and task independent morph decomposition. For all methods LMs for lexicons of three sizes were trained: 100k, 200k and 400k. Going larger than 400k did not show WER improvement in development, most likely due to increase in erroneous words.

The **post-process** method compounds outputs of word-based recognition with a 6m compound replacement dictionary built from LM training data. Two subsequent words are replaced by their compound if the following criteria are met: 1) the words are longer than 3 letters, 2) the words are not very rare, 3) the unigram count of the compound is higher than the bigram count of the words. Alternation of capitalization, insertion of linking 's' and linking hyphens are allowed.

The **decompound** method splits compounds in LM training data into words as in [15] and then post-processes recognition outputs similar to the post-process method. The differences in post-processing are that 1) compounding criteria only use unigram counts, 2) the replacement dictionary was limited to 850k words and 3) pronunciations of the words and the compound must match.

The **morph** method uses unsupervised morphological segmentation to produce morph units instead of words. The morph segmentations are found by applying the Morfessor Baseline algorithm [16], followed by alignment of pronunciations using an HMM and unsupervised tagging of dependent

TABLE II
OOV-RATES AND PERPLEXITIES ON THE TRAINING, DEVELOPMENT AND EVALUATION DATA

Lex.	OOV-rate%				LM perplexity	
	Morphological modeling				Word-based models	
	None	Post-process	Decomound	Morph	3-gram	5-gram
training data						
100k	3.64%	2.32%	2.28%	1.99%	144.3	103.3
200k	2.19%	1.25%	1.21%	0.82%	149.7	106.2
400k	1.31%	0.66%	0.80%	0.30%	153.7	108.5
development data						
100k	1.33%	0.43%	0.54%	0.48%	124.3	99.8
200k	0.64%	0.17%	0.24%	0.15%	128.6	102.7
400k	0.28%	0.07%	0.16%	0.06%	132.1	105.4
evaluation data						
100k	1.42%	0.91%	0.85%	0.73%	160.2	141.2
200k	0.86%	0.57%	0.49%	0.43%	167.4	147.6
400k	0.49%	0.35%	0.35%	0.23%	173.6	153.0

morphs. Morph-based LMs are trained on the text material after converting all words to tagged morph sequences. Each morph is tagged with a pronunciation variant number and a dependency tag to indicate prefixes and suffixes. This enables reconstruction of words after recognition by simple removal of the tags and attached whitespace.

Table II shows the OOV and perplexity reduction obtained by language models and lexicons trained for the three methods. All three methods show some improvement even on the 400k evaluation data OOV-rate. The combination of large lexicon size, lexicon selection and morph-based models provided an OOV-rate of 0.23% on the evaluation data that is very different to the training and development data. On the development data the corresponding number is 0.06%, and in practice the number is even lower, as this only takes into account words that can be constructed from morphs matching the word exactly in pronunciation.

D. Language Modeling

For each of the 26 text components and for each lexicon type and size, a unigram LM was created and the linear combination of the unigrams minimizing perplexity on the development data was searched using SRILM [17]. Using the obtained weights and the word frequency tables of all 26 components, focused lexicons of requested sizes were generated.

Interpolated Modified Kneser-Ney LM training was performed on the 4 main text components using the optimized lexicon, and perplexity minimization was similarly done to find interpolation weights for linear LM interpolation. LMs were not pruned since pruning invariably had a negative effect on the recognition accuracy in development and LM size was not an issue. 5-grams were trained for the word-based models. In morph-based LMs 6-grams were used to compensate for reduced LM span as suggested in [18].

V. RESULTS AND DISCUSSION

Recognition experiments were done on both the development and evaluation datasets. Experiments showing real-time factors were done on a single core 2.2Ghz Opteron 175

TABLE III
WER ON THE DEVELOPMENT AND EVALUATION DATA WITH MORPHOLOGICAL MODELING

Lex.	Morphological modeling			
	None	Post-process	Decomound	Morph
development data				
100k	9.06%	8.41%	8.45%	7.97%
200k	7.90%	7.58%	8.22%	7.38%
400k	7.41%	7.23%	8.43%	7.29%
evaluation data				
100k	21.31%	21.08%	20.87%	20.63%
200k	20.65%	20.55%	21.06%	20.48%
400k	20.32%	20.30%	20.70%	20.42%

processor. Table III lists the WER using the morphological modeling methods on the development and evaluation data. As can be seen, all of the methods give some improvement with the smallest lexicon size, with the gain reducing on the evaluation data and larger lexica.

While the improvement for morph-based LMs in OOV-rate resulted in only minor changes in WER and only for the smaller lexicon sizes, there are several areas to improve and the experiments prove the feasibility of the approach in languages where pronunciations have to be taken into account in modeling morphs. The errors given by the morph-based LMs are also very complementary to that of word-based LMs for purposes of system combination.

Some methods synergistic with morph-based recognition could be integrated to the decoder, such as phone duration modeling, improved pruning and look-ahead language modeling. In addition there is room for improvement in the morph models: morph segmentations from Morfessor have high precision but low recall for large word lists such as here and pronunciation modeling and alignment should be improved, as dense pronunciation lattices complicate morph modeling.

SPRAAK decoder is capable of both efficiently integrating long-span LMs in a single pass and creating word and phone graphs for rescoring in a second pass. Both decoding strategies are popular in LVCSR, and therefore we wanted to see whether two-pass decoding with lattice rescoring is more efficient than one-pass decoding when large long-span n-gram models are available. In theory using a low-order n-gram such as a 3-gram results in a smaller search space, but efficient representation in modern decoders such as SPRAAK reduces this advantage. On the other hand long-span n-gram models can be integrated efficiently, and this gives better LM probabilities for pruning and earlier rejection of incorrect hypotheses.

Table IV shows the results with different configurations on the evaluation and development data, comparing one-pass vs. two-pass decoding, lexicon size, n-gram order and real-time factor. The post-processing method was used and separate results are given for the wideband and telephone segments of development data. The substantially larger WER on the evaluation data is due to a larger portion of the data being spontaneous speech or accented speech. The former is reflected in the high perplexity as seen in table II.

TABLE IV
WER WITH DIFFERENT CONFIGURATIONS FOR DECODING STRATEGY, LEXICON SIZE, N-GRAM ORDER AND REAL-TIME FACTOR

Decoding (LM)		Development, wideband				Development, telephone				Evaluation			
		xRT	WER			xRT	WER			xRT	WER		
pass1	pass2		100k	200k	400k		100k	200k	400k		100k	200k	400k
3-gram	/	0.9	7.88%	6.99%	6.51%	2.3	30.5%	30.0%	30.1%	1.5	23.4%	22.6%	22.8%
	5-gram	1.1	7.05%	6.38%	5.91%	2.7	28.5%	28.0%	28.5%	1.9	21.9%	21.4%	21.2%
3-gram	/	9.0	6.91%	6.01%	5.65%	45.0	27.6%	27.4%	27.0%	37.6	22.5%	21.6%	21.6%
	5-gram	10.5	6.44%	5.65%	5.21%	52.6	25.4%	23.8%	23.6%	57.5	21.1%	20.5%	20.3%
5-gram	/	0.9	7.10%	6.53%	6.18%	2.3	30.2%	30.1%	29.5%	1.5	22.1%	21.8%	21.8%
	5-gram	1.1	7.03%	6.41%	6.05%	2.7	29.6%	29.5%	29.3%	1.9	21.5%	21.2%	21.1%
5-gram	/	2.7	6.69%	5.98%	5.64%	8.0	28.1%	28.3%	28.0%	9.4	21.3%	20.8%	20.7%
	5-gram	3.3	6.67%	5.89%	5.50%	9.9	26.5%	24.6%	26.5%	11.8	21.1%	20.6%	20.4%
5-gram	/	9.0	6.45%	5.67%	5.23%	45.0	26.6%	25.5%	25.8%	37.6	21.1%	20.6%	20.3%
	5-gram	10.5	6.47%	5.65%	5.22%	52.3	25.3%	23.7%	23.3%	49.0	21.0%	20.5%	20.3%

Rescoring with two-pass decoding was most useful with close to real-time systems with low-order n-gram models. However, with higher order n-grams the advantages become considerably reduced, and when a 5-gram is used in the first pass there seems to be no intersection where two-pass decoding would give both lower WER and xRT than one-pass decoding. In addition one-pass decoding avoids a 17% overhead in time use for creating the word graph for rescoring.

Increasing lexicon size almost invariably led to an improvement in WER with the same xRT, the only exceptions found in the telephone segments of development data. These are most likely caused by the lack of sufficient amount of text data for constructing a focused lexicon with interpolation, as with unfocused lexicons larger lexicon sizes lead to inclusion of increasingly more irrelevant words and n-grams. No advantage could be seen from using 3-grams in any condition compared to 5-grams. Rather, with the same xRT roughly 5% relative WER reductions can be seen across the conditions with 5-grams in one-pass decoding.

VI. CONCLUSIONS

This paper described the 2008 ESAT BN transcription system for the Southern Dutch subtask of N-Best. The focus of development was reducing OOV-rate in Dutch and the related issues of language modeling and decoding. We showed that in Dutch the OOV-rate can be reduced to fractions using a combination of structured text normalization, methods for morphological modeling and lexicon selection and efficient search networks to support large lexicon sizes. With token passing decoders such as SPRAAK one-pass decoding with large lexicons and language models is in general preferred over decoding with smaller models. Without model adaptation one-pass decoding was shown to give as good or better results than word graph rescoring.

Comparing to the N-Best evaluation system from LIMS1 [19], their system achieves 8.7% WER on the development data compared to our 7.2%. It is likely that on the evaluation data their system fared better, as their acoustic models integrated all of the available training data and had several adaptation methods. On the other hand on LMs our system is possibly more refined, as disregarding text normalization

differences our LMs showed less than half the perplexity on the development set.

The 2008 N-Best evaluation proved to be a challenging benchmark for development of a recognition system. As a full system for Dutch BN recognition the system presented in this paper is well suited and intended for development and evaluation. For future research on Dutch we plan to both evaluate more methods that are becoming standard in LVCSR and solve the problems specific to Dutch language such as variation in pronunciations. Efficient use of more data should also be able to improve results, as only a very small amount was used for model training here by current standards.

ACKNOWLEDGMENTS

The SPRAAK and N-Best projects were carried out within the STEVIN program funded by the Dutch and Flemish governments (<http://www.stevin-tst.org/english>). This research was also supported by the Fund for Scientific Research Flanders (FWO Project "TELEX" G.0260.07).

REFERENCES

- [1] J. Kessens and D. A. v. Leeuwen, "N-best: the Northern- and Southern-Dutch benchmark evaluation of speech recognition technology," in *Proc. of INTERSPEECH*, 2007, pp. 1354–1357.
- [2] T. L. D'Halleweyn E., Odijk J. and C. C., "The Dutch-Flemish HLT programme STEVIN: Essential speech and language technology resources," in *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, May 2006, pp. 761–766.
- [3] K. Demuyneck, J. Roelens, D. Van Compernelle, and P. Wambacq, "SPRAAK: An open source speech recognition and automatic annotation kit," in *Proc. ICSLP*, Brisbane, Sep. 2008, p. 495.
- [4] N. Oostdijk, "Het Corpus Gesproken Nederlands," *Nederlandse Taalkunde*, vol. 5, no. 3, pp. 280–284, 2000.
- [5] K. Demuyneck, "Extracting, modelling and combining information in speech recognition," Ph.D. dissertation, K.U.Leuven, Feb. 2001.
- [6] J. Duchateau, M. Wigham, K. Demuyneck, and H. Van hamme, "A flexible recogniser architecture in a reading tutor for children," in *Proc. of ITRW on Speech Recognition and Intrinsic Variation*, Toulouse, France, May 2006, pp. 59–64.
- [7] P. Mertens and F. Vercammen, "FONILEX manual," K.U.Leuven – CCL, Technical report, 1998.
- [8] J. Duchateau, "HMM based acoustic modelling in large vocabulary speech recognition," Ph.D. dissertation, K.U.Leuven, Nov. 1998.
- [9] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems 14*. MIT Press, 2001, pp. 849–856.
- [10] K. Demuyneck, J. Duchateau, and D. Van Compernelle, "A static lexicon network representation for cross-word context dependent phones," in *Proc. EUROSPEECH*, vol. I, Rhodes, Greece, Sep. 1997, pp. 143–146.

- [11] K. Demuynck, J. Duchateau, and D. V. Compernelle, "Reduced semi-continuous models for large vocabulary continuous speech recognition in Dutch," in *Proc. of International Conference on Spoken Language Processing, volume IV*, 1996, pp. 2289–2292.
- [12] M. Hwang, X. Huang, and F. Alleva, "Predicting unseen triphones with senones," in *Proceedings of ICASSP*, 1993, pp. 311–314.
- [13] K. Demuynck, T. Laureys, P. Wambacq, and D. Van Compernelle, "Automatic phonemic labeling and segmentation of spoken Dutch," in *Proc. LREC-2004*, Lisbon, Portugal, May 2004, pp. 61–64.
- [14] J. R. Quinlan, *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.
- [15] B. Réveil and J.-P. Martens, "Reducing speech recognition time and memory use by means of compound (de-)composition," in *Proceedings ProRisc 2008*, Veldhoven, The Netherlands, Nov. 2008, pp. 348–352.
- [16] M. Creutz and K. Lagus, "Inducing the morphological lexicon of a natural language from unannotated text," in *Proc. of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, Espoo, Finland, Jun. 2005, pp. 106–113.
- [17] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. of the International Conference on Spoken Language Processing*, 2002, pp. 901–904.
- [18] T. Hirsimäki, J. Pytkönen, and M. Kurimo, "Importance of high-order n-gram models in morph-based speech recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 17, no. 4, pp. 724–732, 2009.
- [19] J. Despres, P. Fousek, J.-L. Gauvain, S. Gay, Y. Josse, L. Lamel, and A. Messaoudi, "The joint LIMSI and Vecsys research systems for NBEST 2008," Sep. 2008, presented in N-Best Workshop.