INTERSPEECH
2009 BRIGHTON

Modeling Northern and Southern Varieties of Dutch for STT †

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, CNRS-LIMSI, BP 133, 91403 Orsay cedex, France
[2] Vecsys Research, 3, rue Jean Rostand, 91400 Orsay, France
{gauvain,lamel,fousek,abdel}@limsi.fr, {despres,gay,josse}@vecsysresearch.com

Abstract

This paper describes how the Northern (NL) and Southern (VL) varieties of Dutch are modeled in the joint LIMSI-Vecsys Research speech-to-text transcription systems for broadcast news (BN) and conversational telephone speech (CTS). Using the Spoken Dutch Corpus resources (CGN), systems were developed and evaluated in the 2008 N-Best benchmark. Modeling techniques that are used in our systems for other languages were found to be effective for the Dutch language, however it was also found to be important to have acoustic and language models, and statistical pronunciation generation rules adapted to each variety. This was in particular true for the MLP features which were only effective when trained separately for Dutch and Flemish. The joint submissions obtained the lowest WERs in the benchmark by a significant margin.

Index Terms: speech recognition, Dutch, Flemish, CGN, N-best, broadcast news, conversational telephone speech, MLP.

1. Introduction

This paper describes the speech-to-text transcription systems developed at LIMSI and Vecsys Research to process Broadcast News (BN) and Conversational Telephone Speech (CTS) data in two main varieties of Dutch (Northern and Southern) as spoken by people from The Netherlands and from Flanders (Belgium), respectively. It was found to be beneficial to specifically model the two varieties in both the acoustic and language models. Concerning lexical modeling, separate grapheme-to-phoneme systems (sharing the same phone set) were used for the two varieties, and these were merged to form the final common lexicon in order to share the pronunciation probabilities.

The 2008 N-Best (Northern and Southern Dutch Benchmark Evaluation of Speech recognition Technology) project of the Dutch-Flemish Stevin program (speech.tm.tno.nl/n-best) organized a benchmark evaluation in large vocabulary speech recognition for the Dutch language. The evaluation was conducted by TNO Human Factors Soesterberg, the Netherlands in cooperation with Spex in Nijmegen, and aims to foster the development of speech corpora and technologies for the Dutch language [13]. The participants in the benchmark were provided with 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	1.1 M	572.2 K
CTS	92.0 / 80.0	64.0 / 60.0	1.3 M	808.3 K

Table 1: *N-Best acoustic data provided by CGN (total data / transcribed training data).*

2. Task and data description

The baseline acoustic and language modeling training data are shown in Table 1 according to variety and type. There are about 100 hours of audio data for Northern Dutch (NL) and over 50 hours for Southern Dutch (VL), with (∼1.2M words) and (∼700K words) of manual transcripts respectively. Since the development data sets were also taken from the CGN data, they had to be removed from training. The language model training data are comprised newspaper articles from 1999 to 2004, obtained from the Dutch publisher PCM and the Flemish Mediargus. The data contain approximately 360M words of Dutch and 1418M words of Flemish, with respectively about 7.2M and 14.8M distinct lexical forms. Table 2 summarizes the development data. For the CTS audio files the conversations were recorded on two different channels, each channel was decoded separately. The development files were mainly composed of CGN excerpts, except for the BN-NL task which also included parts from another data source. There is a total of about 1 hour of speech (∼9K words) for each BN task and just under 2 hours (∼7K words) for each CTS task.

3. Speech recognizer overview

This section gives an overview of speech recognizers used in this work, more details about the models are given in Sections 4-6. The recognizers use the same basic statistical modeling techniques and decoding strategy as in the LIMSI English broadcast news system [8]. Prior to transcription, an audio partitioner divides the continuous audio stream into homogeneous segments, associating cluster, gender and labels with each non-overlapping segment [7, 8]. For the CTS the clustering step is not needed since all speech segments are assumed to come from the same speaker. The acoustic and language models are language and task specific. The Dutch and Flemish lexicons use the same word list, phone symbol set and pronunciation variants, but the pronunciation probabilities collected during the acoustic training are task-specific.

The primary recognition submission results from a ROVER [4] between two system outputs, using different acoustic features: PLP and MLP, each generated in 2 decoding passes. Each of these systems include rescoring by a 4-gram neural net-

† This work was in part supported by OSEO under the Quaero program.

Task	*Data*	*Total*	*Scored*	
	Type	*Duration (h)*	*Duration (h)*	*#words*
BN-NL	Dev.	1.1	1.0	8721
BN-VL	Dev.	1.0	1.0	10406
CTS-NL	Dev.	2.0 (x2 ch.)	1.8	6695
CTS-VL	Dev.	1.9 (x2 ch.)	1.8	6790

Table 2: *Development data. Scored duration corresponds to the duration of the segments given by the UEM file (only these segments, whose size in words is given, are scored).*

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 3: *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.*

work 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 systems process the audio data in under 10 times real-time. The 1xRT word recognition is performed in a single decoding pass, using a 2-gram LM for decoding and a 4-gram LM for rescoring. Table 3 summarizes the submissions.

4. 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) [6, 20]. 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 Dutch and Flemish MLPs were trained for each task [1], 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.

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

Table 4: *Recognition lexicons. For each word list, separate lexicons are generated for each variety, and the two are merged.*

All acoustic models (AMs) are tied-state, left-to-right context-dependent (CD), HMMs with Gaussian mixtures. The triphone-based CD phone models are word-independent but position-dependent. The tied states are obtained by means of a decision tree. Different sets of gender-independent AMs were trained for each task (BN and CTS), and each variety (NL and VL). The models all use speaker-adaptive (SAT) and Maximum Mutual Information Estimation (MMIE) training. For each task and variety, 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.

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 speaker and region-independent models are trained on all of the available data for the task (130 hours for BN and 150 hours for CTS). These models serve as priors for Maximum *a Posteriori* (MAP) [10] estimation of variety-specific models for each task.

5. Pronunciation lexicons

All pronunciations are based on a set of 41 phones (16 vowels, 22 consonants and 3 other symbols that represent silence, filler words, and breath noises). These phones 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.

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 [16] dictionary.

The pronunciation lexicon is formed by associating a list words (see Section 6) with one or more pronunciations. If the words are present in the Dutch/Flemish master lexicon their pronunciations are extracted. Words for which no pronunciation is present in the master lexicon are phoneticized by a statistical approach using the translation tools Giza++ [17] and Moses [12]. This approach was inspired by the method described by Dealemans and 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 summa-

[1] Features produced by a single MLP trained on both varieties were less effective.

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 5: *Language model development. All models were generated using a cut-off of 1-2-3 and a pruning value of 1e-10.*

rized in Table 4. 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) obtained via forced alignment of the training data with their orthographic transcriptions.

6. Language modeling

To facilitate training, common word lists were used for Dutch and Flemish including all words in the audio training transcripts and the most frequent words in the text corpora.[2] The vocabulary size (*n*) was chosen to minimize the OOV rate on the 4 development data sets. With a 300K case-sensitive word list, the OOV rate is under 1% for all 4 data sets. The OOV rate is further reduced to about 0.5% with a 500K case-sensitive word list.

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 [19]) trained on separate subsets of the available language model training texts. The characteristics of the language models are summarized in Table 5. 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 texts (10 sources): 357M words (cut-off 0-1-2)
3) VL texts (12 sources): 1215M words (cut-off 0-1-2)

AM Training	*BN-NL*	*BN-VL*
NL (84h) / VL (48h)	21.7	24.2
NL+VL (132h)	20.5	23.0
NL+VL ⇒ NL/VL	20.2	21.2

Table 6: *Case sensitive WER (%) on dev08 with different acoustic model training (NL or VL only, pooled, pooled+MAP) and with 87k word NL or VL specific LMs estimated only on the training transcripts.*

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 7: *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.*

The mixture weights were automatically chosen by an EM algorithm to minimize the perplexity of the development data. The 2-gram 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.

7. Experimental results

Initial model development was carried out using only the BN audio data and associated transcriptions. The first acoustic models were trained separately for each variety and used in a first pass decode for that variety, The initial case-sensitive word error rates (WER) are shown in the first entry of Table 6. Pooling together all the audio is seen to improve both varieties (NL+VL), and an additional gain is obtained using MAP adaptation for each one, with a notably larger improvement for VL for which there is less training data. (Using cross-variety acoustic models degrades performance by about 20% relative.) All further acoustic model development used the pooled data with variety adaptation. The next series of experiments were directed at improving the language models, exploring different text normalizations (mainly affecting the definition of a word) and using the newspaper training texts. System development was primarily carried out for the BN task, after which the same strategies were applied for CTS.

For the final system, 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 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 [14]. The lattices produced in the last pass 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:

[2] There are about 66 K/44 K distinct lexical items in the NL/VL audio transcripts, with only 23 K common words. For the significantly larger LM texts, a similar proportion of the distinct words were also shared.

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 8: *Final case-sensitive/case-insensitive WER (in %) on the dev08 data for the 4 tasks for the primary (also under 10xRT) system and the contrastive systems (1xRT: PLP-pass1). Punctuation and non-lexical events are not scored.*

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

1) Initial hypothesis generation using large MLLT and MMIE-trained AMs (∼ 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 7 and 8 give the word error rates on the N-Best 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, the WER decreases by about 1% on average for the BN tasks, but only by about 0.2% on the CTS tasks.

Table 9 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 9 to the pass 2 results in Table 7). For the MLP based system, a slower second pass decoding is carried out, which results in a WER reduction 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 8 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.

8. Summary

This paper has given an overview of the way in which Northern and Southern variety and task-specific models were developed for the speech recognizers that served as a joint submission by LIMSI and Vecsys Research to the Dutch N-Best 2008 evaluation. In total 8 systems were developed (1xRT and 10xRT), for each variety (NL and VL) and task (BN and CTS). It was found that techniques used for other languages were generally also effective for Dutch. Given the differences between the two varieties, using variety-specific models was also found to be important. At the acoustic level, variety-specific models were obtained by MAP adaptation of speaker-independent models trained on all the data for the task. Variety-specific pronunciation rules and dictionaries were used in the early stages of system development, and then merged to simplify the data sharing. The statistical method made use of translation tools to generate multiple pronunciations for words not present in the available dictionaries. A common word list was used for all tasks and varieties, with different language model interpolation coefficients for the different conditions. Word error rates under 10% were obtained on BN development data, and on the order of 30% for Dutch and 40% Flemish conversational data. On the evaluation data, these systems obtained the lowest word error rates (17.8% BN-NL, 15.9% BN-V, 35.1% CTS-NL and 46.1 CTS-VL) by a significant margin.

9. References

[1] Baayen, R.H., Piepenbrock, R., Gulikers, L., ”The CELEX lexical database”. Linguistic Data Consortium: LDC96L14, 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., van den Bosch, A., ”A language-independent, data-oriented architecture for grapheme-to-phoneme conversion”, *ESCA-IEEE conference on Speech Synthesis*, NY, 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, 2007.

[6] Fousek, P., Lamel, L., Gauvain, J.L., ”Transcribing Broadcast Data Using MLP Features,” *ICSLP'08*, 1433-1436, 2008.

[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., 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., 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 & Audio Processing*, **2**(2):291-298, 1994.

[11] Grézl, F., Fousek, P., ”Optimizing bottle-neck features for LVCSR,” *IEEE ICASSP'08*, Las Vegas, 2008.

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

[13] D.vanLeeuwen et al, “Rsults of the N-Best 2008 Dutch Speech Recognition Evaluation”, these proceedings.

[14] 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.

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

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

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

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

[19] Stolcke, A., ”SRILM – an extensible language modeling toolkit,” 2002.

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