

N-best: The Northern- and Southern-Dutch Benchmark Evaluation of Speech recognition Technology

Judith Kessens and David van Leeuwen

TNO Human Factors, Soesterberg, The Netherlands

judith.kessens@tno.nl, david.vanleeuwen@tno.nl

Abstract

In this paper, we describe N-best 2008, the first Large Vocabulary Speech Recognition (LVCSR) benchmark evaluation held for the Dutch language. Both the accent as spoken in the Netherlands (Northern-Dutch) and in Belgium (Southern-Dutch or Flemish), will be evaluated. The evaluation tasks are broadcast news (BN) and conversational telephone speech (CTS). The N-best evaluation will take place in the spring of 2008 and is open to all research institutes and industries on voluntary basis. The goals of this first N-best evaluation is to define, set-up and conduct a Dutch LVCSR benchmark evaluation. In this paper, we will describe the state-of-the-art of Dutch LVCSR, recognition problems that are typical for the Dutch language, and the evaluation protocol.

Index Terms: Northern- and Southern-Dutch, large vocabulary speech recognition, benchmark test, evaluation, conversational telephone speech, broadcast news.

1. Introduction

Evaluations in speech technology research allow comparison of approaches and are an incentive to speech researchers to develop new techniques. An evaluation of speech technology is both useful as a benchmark for the present, and for developing new approaches in the future. With the completion of the ‘Spoken Dutch Corpus’ (CGN, [1]) most prerequisites for the development of Dutch LVCSR systems are in place, but thus far a proper evaluation has not been defined or held. The lack of a proper evaluation for Dutch has often led researchers to do research using foreign languages, such as English. The N-best project aims at setting up the infrastructure for a benchmark evaluation in large vocabulary speech recognition for the Dutch language, and at conducting such an evaluation.

The N-best evaluation will be conducted along the lines of other evaluations in speech technology as the well known series of evaluations by the National Institute of Standards and Technology (NIST, [2]) in the US, the EU SQALE [3] project and the French *Technolangue* evaluation series (see e.g. [4]).

In this paper, we will describe the state-of-the-art of Dutch LVCSR (section 2), recognition problems that are typical for the Dutch language (section 3) and the N-best 2008 evaluation protocol (section 4).

2. State-of-the-art of Dutch LVCSR

In literature, recognition results on LVCSR in Dutch are not often reported: Word Error Rates (WERs) can be found for

three broadcast news LVCSR systems developed by the University of Twente (ABBOT, Sonic, SHoUT) and one spontaneous speech LVCSR system, developed by the University of Leuven (ESAT).

Ordelman [5] has ported the ABBOT system to Northern-Dutch. This system is based on hybrid RNN/HMM acoustic models, a 65K vocabulary and a statistical trigram language model, based on a news corpus. On the Twente News Corpus (TwNC), consisting of 10 TV news shows, a WER of 32% is obtained [5]. Later, the same system was tested by Huijbregts et al. [6] on a 4 hrs test set consisting of BN selected from the CGN corpus, resulting in a WER of 35%. On the same CGN test set, a second system was tested. For this system, based on the Sonic speech recognizer [7], a lower WER of 30% is reported. The system uses decision-tree state-clustered Hidden Markov Models. Twenty-two hours of broadcast news recordings from the CGN were used to port the English acoustic models to Dutch. The same language-model and vocabulary is used in both tests. Finally, the University of Twente developed the SHoUT system, described in [8]. SHoUT has been trained on 79 hrs of BN, interviews and live commentaries of sport events from the CGN corpus. BN language models (LMs) were estimated from approximately 500 million words of normalised Dutch text data from various resources (mainly newspaper data). The decoder's Viterbi search is implemented using the token passing paradigm. HMMs with three states and GMMs for its probability density functions are used to calculate acoustical likelihoods of context dependent phones. Trigram backoff language models (LMs) are used to calculate priors. SHoUT has been evaluated on one BN show from the TwNC and on BN data from the CGN. With Vocal Tract Length Normalisation, WERs of 29% and 19% were obtained on the TwNC and CGN test sets, respectively.

For Southern-Dutch, an LVCSR system has been developed by ESAT in Leuven [9]. The acoustic models are trained on 44 hrs of Southern-Dutch, spontaneous speech from the CGN corpus. About 3500 tied states model the acoustics of context-dependent phones. The system has been tested on 5k word test set selected from the Southern-Dutch spontaneous broadcast data [10]. The ESAT decoder performed a single pass time synchronous beam search, which results in real-time recognition. A 40k recognition vocabulary was used, and various language models have been tested. The best language model was trained on 40.5M words from the CGN spontaneous speech transcriptions and a selection of newspaper texts. In another study [11] the ESAT LVCSR was used to recognize a selection of spontaneous speech from the CGN database (interviews, field reports, debates, discussions). To this end, the 40k recognition vocabulary was supplemented in order to attain full lexical coverage. A WER

of 36% is reported on the spontaneous speech recognition task.

A summary of the WERs on Dutch LVCSR is given in Table 1. Although these performance results can not be directly compared to those of English LVCSR systems, it can be concluded that the reported error rates do not compete with international state-of-the-art English systems: for the DARPA EARS rich transcription program, the best recognition output in the RT04F evaluation was 11.7% for BN and 14.9% for CTS. In the RT03S evaluation, an even lower WER (under 10%) was achieved for BN using a system that was not time limited.

A couple of commercial LVCSR speech recognizers are available: The Nuance product ‘Audio Mining’ is supported by an LVCSR system trained on BN and telephone speech. Autonomy has a Softsound speech-to-text engine for processing Dutch BN. However, for all these products, no performance measures are reported for Dutch.

These observations support the need for a Dutch LVCSR benchmark evaluation infrastructure, exploiting and sharing the various data collection efforts. Furthermore, there seem to be enough possibilities for improving current state-of-the-art LVCSR systems.

Table 1. Recognition results on Dutch LVCSR

ref	system	test set	dial.	d-base	WER
[5]	ABBOT	news shows	ND	TwNc	32%
[6]		broadcast news	ND	CGN	35%
	Sonic				30%
[8]	SHoUT	one news show	ND	TwNc	29%
		broadcast news		CGN	19%
[10] [11]	ESAT	spontaneous speech	SD	CGN	34-36%

3. Dutch speech recognition

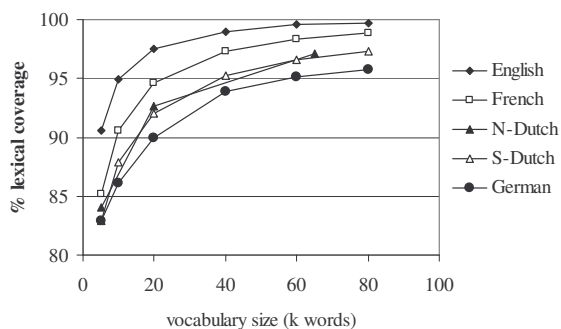


Figure 1: Lexical coverage for Northern- and Southern-Dutch and three other languages

One of the goals of the N-best evaluation is to share research effort on typical problems of Dutch speech recognition. One typical recognition problem is compounding and word inflections, similar to German. Compared to English, the Dutch language has a high number of compounds and word inflections. Therefore, a larger lexicon is needed in Dutch to achieve the same lexical coverage as in English. Figure 1 illustrates the differences in lexical coverage among four

languages. The lexical coverages for English, French and German are adapted from [3], for Northern-Dutch from [13], and for Southern-Dutch from [17].

Besides compounding, language-related issues can be investigated within the N-best framework, e.g. optimizing the Dutch phone set, choosing decision tree questions and handling Dutch pronunciation variation.

4. The N-best evaluation protocol

The N-best 2008 evaluation will be co-ordinated by TNO Human Factors in the Netherlands (evaluator), and is open to all research institutes and industries on voluntary basis. Organizations participating in the evaluation will be named ‘ASR sites’ hereafter.

In order to advise the evaluator in setting up the evaluation protocol, a working group has been compiled consisting of: the Radboud University (CLST), Delft University of Technology (EWI), University of Twente (HMI) in the Netherlands, and the Universities of Gent (ELIS) and Leuven (ESAT) in Belgium. All partners of the working group will be participating in the N-best 2008 evaluation. The project therefore aims at filling in the benchmark results for at least four LVCSR technologies (ESAT-FLaVoR [19], ESAT-SPRAAK [18], HTK [20], Abbot [21] and Sonic [7]).

4.1. Tasks

The evaluation tasks are conversational telephone speech and broadcast news. These tasks have been chosen for three reasons: 1. these tasks are best known in the international benchmarks in speech recognition, and will allow for cross language comparison, 2. for both tasks acoustic and textual training data are available, and 3. the tasks have a strong application potential.

Four primary evaluation tasks are defined by the two tasks (CTS and BN) and two dialects (Northern- and Southern-Dutch). Results will be analyzed separately for these four tasks. Although a site can decide to optimize on only one of the four tasks, all tasks have to be performed to obtain a valid submission.

4.2. Conditions

Several aspects of the speech recognition task will be conditioned. *Primary* conditions refer to conditions that have to be performed minimally for valid submission. Primary conditions are defined to measure baseline recognition performance and for comparison of the various systems. In the primary condition, no limit is imposed on the processing time. Furthermore, to train the acoustic and language models only a predefined acoustic and a textual database may be used.

Contrastive conditions may be optionally performed, additionally to the primary condition. Some contrastive conditions are pre-defined, but a site may wish also to define its own contrastive condition(s). Contrastive conditions are meant to stimulate researchers to test other aspects than the primary baseline conditions. Two contrastive conditions have been pre-defined for processing speed: real time (RT) and 10x RT.

The choices for, e.g., the vocabulary, pronunciation dictionary, and phone set are not constrained. These choices are considered design choices and are part of the differences between various systems under evaluation. However, the training material will be delivered with a pronunciation dictionary that covers the training data. Although this information is not distributed by the evaluator, sharing information between sites or defining new contrastive conditions is encouraged.

4.3. Materials

4.3.1. Evaluation material

The evaluation material is recorded by SPEX [22] after a specific date, namely January 1st 2007. For each of the four primary tasks, two hours of evaluation material will be recorded.

For the BN, the recordings will be made from public and commercial radio and television broadcast in the Netherlands and Belgium. The BN material will consist of about 10 excerpts from 10 different news shows. Each excerpt consists of fragments taken from a single show. We choose relatively long excerpts to make it possible to perform speaker clustering/adaptation within one news show. The BN material will be chosen from a wide range of news shows and broadcast channels. Segmentation of the BN audio file in speech, music, jingle and other audio types is not an item under evaluation. Therefore, audio excerpts will be chosen that contain primarily speech. The BN-audio will not be segmented at the sentence level, but longer segments will span several speakers and acoustical conditions.

For CTS, spontaneous telephone dialogs will be recorded with Northern- or Southern-Dutch speakers. The conversation protocol is similar to the Switchboard protocol [23]. Subjects talk about a conversation topic, chosen from a list of topics. The CTS evaluation material will consist of two-channel files, for which both sides of the conversation have to be processed. The segments will include the ‘silent’ parts of conversations sides. The material will originate from 12 dialogues of around 10 minutes, where each conversation side will speak roughly half of the time. The CTS evaluation material will be balanced as much as possible for important speaker characteristics (e.g. age, sex, region).

4.3.2. Training material

Acoustical and language model training material will be specified and can be provided to participating ASR sites. In the primary condition, this material can be used for building Dutch acoustic and language models. The acoustic material will be a selection of the ‘Corpus Gesproken Nederlands’ (CGN) [1]. The CGN contains both Northern- and Southern-Dutch speech material. For CTS, a selection of CGN components c+d will be made, whereas for the BN, components f+i+j+k+l will be used for selection of the training material [25]. For each speech domain and dialect, between 53 and 99 hours of speech material (incl. silences) will be available.

The training material for the language model will be obtained from two resources: the Dutch publisher PCM (360 million words) and the Flemish Mediargus (1,436 million words). From both resources, a collection of newspapers over a fixed period will be chosen.

4.3.3. Development material

Part of the acoustic training data will be split off as development data. This development data will be used for performing a dry-run of the evaluation. The development material will be distributed in the same way as the evaluation material. The selection of the development data is done in such a way that it resembles the evaluation data (type of speech, number of speakers, length of speech fragment, type of show, ratio between radio/television material). Splitting will be performed in such a way that the development data will originate from a more recent period than the training data.

4.4. Evaluation measure

The primary evaluation measure will be Word Error Rate (WER), as calculated by NIST sclite tools [1]. Alignment between the reference transcription and ASR output is carried out in such a way that the WER is minimized. The WER values will be reported for all four primary tasks separately. Non-lexical events (coughs, hesitations, filled pauses) will not be included in the primary WER scores but will be analyzed separately.

In contrast to what is customary in evaluations in English, the WER scoring will be case-sensitive. Although in the SQALE-project [3] it has been shown that removing case sensitivity for French and German resulted in only 0.1-0.3% drop in error rate, we will investigate whether this is the case for Dutch as well. First words of sentences will not be capitalized. Punctuation is not an item under evaluation.

The WER will also be calculated in a time-mediated way, and will be reported separately. For sites that include word-based confidence measures, these confidence measures will be evaluated along similar lines as NIST Normalized Cross Entropy [16].

Especially for the broadcast news task, the acoustic conditions will be variable. By coding the acoustic conditions according to the NIST focus conditions [24], we will be able to compare the results per acoustic condition with the NIST benchmark results.

4.5. Time schedule

For the organization of the evaluation, we followed the organization of almost all (DARPA) NIST evaluation campaigns and as recommended by the ELSE project [14]:

1. Training,
2. Dry run,
3. Evaluation,
4. Impact study.

During the training phase the systems will be trained and optimized. The second phase consists of an optional ‘dry-run’, where the evaluation will be ‘simulated’ using the development database. Next, the evaluation will take place, meaning that after distribution of the evaluation material, the output of the ASR-system must be produced within a fixed period (one month). Next, the evaluator will score the results after which there will be an adjudication period in which ASR sites will be given the opportunity to comment on certain interpretations and decisions made by the evaluator. Finally, in a workshop, the evaluator and ASR sites will present their results to the others. During the workshop TNO will present

the overall results and will try to draw general conclusions based on comparison of the results obtained. The important dates are summarized in Table 2.

Table 2. *Global time schedule of N-best 2008.*

Date	Activity
jan. '07	Start of collection of evaluation material
sept. '07	Dry-run
apr. '08	Evaluation.
may '08	Adjudication period
june '08	One-day workshop at TNO

4.6. Dissemination

In order to facilitate current and future researchers to use the benchmark evaluation results we will adhere to international and open standards for file formats and evaluation measures. In practice, this means that wherever possible we will make use of the evaluation structure and software that organizations such as NIST (US) and DGA (France) have defined. Any information that is exchanged will be in open format files and media. After the evaluation has ended, the training, development and evaluation databases and the evaluation protocol will be made available via the 'TST-centrale' (distributor of Dutch language and speech resources [26]).

5. Acknowledgement

This N-best 2008 evaluation is supported by the STEVIN programme. STEVIN aims at stimulating Dutch language and speech technology and realizing a digital infrastructure for Dutch language- and speech material essential for developing these technologies. More information on the N-best project and the N-best 2008 evaluation can be found at: <http://speech.tn.tno.nl/n-best>.

6. References

[1] Oostdijk, N.H.J., Broeder, D., 'The Spoken Dutch Corpus and its exploitation environment', *Proc. of LINC-03*, Budapest, Hungary, 2003.

[2] Pallett, D., "A look at NIST's Benchmark ASR Tests: Past, Present, and Future", www.nist.gov/speech/history/pdf/NIST_benchmark_AS_Rtests_2003.pdf, 2003.

[3] Young, S., Adda-Decker, M., Aubert, X., Dugast, C., Gauvain, J.-L., Kershaw, D.J., Lamel, L., van Leeuwen, D., Pye, D., Robinson, A.J., Steeneken, H.J.M., Woodland, P.C., 'Multilingual large vocabulary speech recognition: the European SQALE project', *Computer, Speech and Language*, 11:73-89, 1997.

[4] Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., Gravier, G., 'The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News', *Proc. Interspeech'05*, Lisboa, Portugal, pp. 1149-1152, 2005.

[5] Ordelman, R., 'Dutch Speech Recognition in Multimedia Information Retrieval', *PhD. thesis*, University of Twente, ISBN 90-75296-08-8, 2003.

[6] Huijbregts, M.A.H., Ordelman, R.J.F., and de Jong, F.M.G., 'A Spoken Document Retrieval Application in the Oral History Domain', *Proc. of 10th int. conference Speech and Computer*, University of Patras, pp. 699-702, 2005.

[7] Pellom, B., 'SONIC: The University of Colorado Continuous Speech Recognizer', University of Colorado, tech report #TR-CSLR-2001-01, Boulder, Colorado, March, 2001.

[8] Huijbregts, M., Ordelman, R., de Jong, F., 'Annotation of Heterogeneous Video Content Using ASR', submitted to Interspeech'07.

[9] Demuyneck, K., Duchateau, J., Van Compernelle, D., and Wambacq, P., 'An Efficient Search Space Representation for Large Vocabulary Continuous Speech Recognition', *Speech Communication* (30) 1, pp. 37-53, January 2000.

[10] Duchateau, J., Van Uytsel, D.H., Van hamme, H., and Wambacq, P., 'Statistical Language Models for Large Vocabulary Spontaneous Speech Recognition in Dutch', *Proc. of Eurospeech'05*, Portugal, pp. 1301-1304, 2005.

[11] Stouten, F., Duchateau, J., Martens, J.-P., and Wambacq, P., 'Coping with disfluencies in Spontaneous Speech Recognition: acoustic detection and linguistic context manipulation', *Speech Communication* (48), pp. 1590-1606, 2006.

[12] Liu, Y., Shriberg, E., Stolcke, A., Peskin, B., Ang, J., Hillard, D., Ostendorf, M., Tomalin, M., Woodland, P., Harper, M., 'Structural metadata research in the EARS program', *Proc. of ICASSP'05*, vol. 5, pp. 961-964, 2005.

[13] Ordelman, R.J.F., van Hessen, A.J., and de Jong, F.M.G., 'Compound Decomposition in Dutch Large Vocabulary Speech Recognition', *Proc. of Eurospeech'03*, pp.225-228, 2003.

[14] Paroubek, P. and Blasband, M., 'Evaluations in Language and Speech Engineering', www.limsi.fr/TLP/ELSE/, 1999.

[15] www.nist.gov/speech/tools/index.htm

[16] www.nist.gov/speech/tests/rt/rt2004/fall/docs/NCE.pdf

[17] Laureys, T., Vandeghinste, V., & Duchateau, J., 'A Hybrid Approach to Compounds in LVCSR', *Proc. of ICSLP'02*, Denver, pp. 697-700, 2002.

[18] <http://www.spraak.org/>

[19] Demuyneck, K., Laureys, T., Van Compernelle, D. and Van hamme, H., "FLaVoR: a Flexible Architecture for LVCSR", Proc. European Conference on Speech Communication and Technology, pp. 1973-1976, Geneva, Switzerland, September 2003

[20] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., "The HTK Book", Cambridge University, Cambridge, UK, Vol. 2, pp. 829-832, 2002.

[21] Robinson, T., Christie, J., 'Time-first search for large vocabulary speech recognition', Proc. of ICASSP'98, 1998.

[22] Speech Processing Expertise Centre (SPEX), Nijmegen, the Netherlands, www.speex.nl.

[23] Godfrey, J. J. and Holliman, E. C. and McDaniel, J., 'SWITCHBOARD: telephone speech corpus for research and development', Proc. of ICASSP'92, pp. 517-520, 1992.

[24] Stern, R.M., 'Specification of the 1996 HUB 4 Broadcast News Evaluation', 1997 DARPA Speech Recognition Workshop, February 2-5, Chantilly, Virginia, 1997.

[25] <http://lands.let.kun.nl/cgn/ehome.htm>

[26] <http://www.tst.inl.nl/>