

Results of the N-Best 2008 Dutch Speech Recognition Evaluation

David A. van Leeuwen¹, Judith Kessens¹, Eric Sanders² and Henk van den Heuvel²

¹TNO Human Factors, Soesterberg, The Netherlands

²SPEX, Nijmegen, The Netherlands

Abstract

In this paper we report the results of a Dutch speech recognition system evaluation held in 2008. The evaluation contained material in two domains: Broadcast News (BN) and Conversational Telephone Speech (CTS) and in two main accent regions (Flemish and Dutch). In total 7 sites submitted recognition results to the evaluation, totalling 58 different submissions in the various conditions. Best performances ranged from 15.9% word error rate for BN, Flemish to 46.1% for CTS, Flemish. This evaluation is the first of its kind for the Dutch language.

1. Introduction

In the Flemish-Netherlands Language and Speech Technology programme STEVIN¹ the project N-Best (Northern and Southern Dutch benchmark evaluation for speech recognition technology) aimed at setting up the infrastructure for and running an evaluation for large vocabulary continuous speech recognition systems (LVCSR) in the Dutch language. The project also stimulated several research groups in the Netherlands and Flanders to (further) develop a LVCSR, comprised collecting new evaluation data and provisioned in sharing experiences and resources for training the systems.

The design of the evaluation is very similar to the US DARPA sponsored NIST evaluation campaigns in the 90's and early this century [1], and to the French ESTER evaluation [2] of the Technolangu programme. Since it is the first LVCSR evaluation in Dutch language we decided to start with the well-established speech domains of Broadcast News (BN) and Conversational Telephone Speech (CTS), and not investigate other domains such as Meeting data—which is currently used in the NIST Rich Transcription series of evaluations. The choice of speech domains is also motivated by the availability of training data for Dutch, which in a large part is formed by the Spoken Dutch Corpus (*Corpus Gesproken Nederlands*, CGN) [3], in which the BN and CTS parts may be closest in style to their traditional English counterparts. The Dutch spoken in the low countries can be classified in two broad accent groups, Northern and Southern Dutch, often indicated as Dutch (NL) and Flemish (VL)—corresponding to the countries The Netherlands and Belgium, respectively), a difference that may be reinforced because of social and cultural reasons such as radio and television shows.

The project, led by TNO, was started in 2006, and after preparation and a dry-run test, the evaluation took place in April 2008. Five Automatic Speech Recognition (ASR) sites in the STEVIN N-Best project participated: ELIS (University of Gent), ESAT (KU Leuven), Radboud University Nijmegen, EWI (Delft University of Technology) and HMI (University of

Table 1: Amount of training data form CGN in primary training condition. The CGN components refer to the 15 main speech types in CGN.

Domain	CGN component	NL	VL
BN	f, i, j, k, l	99 h	53 h
CTS	c, d	92 h	64 h

Twente). Two additional sites participated: Vecsys Research + LIMSI (France), and Brno University of Technology (Czech Republic). These 7 sites will be referred to as 'ASR sites' hereafter.

We will not, in this paper, get into details about the technological difference between systems to hypothesize what possible reasons for difference in performance are. Rather, this paper will review the evaluation process and data, and present the main evaluation results.

2. Evaluation task and protocol

The task and rules of evaluation were specified in an evaluation protocol [4]. There were four *primary tasks* defined, formed by the Cartesian product of the speech domains BN and CTS, and the accents NL and VL. The task was that of *transcription* (ASR or speech to text), with the focus on lexical content, i.e., ignoring filler words, hesitations etc. in scoring. Approximately 2 hours of speech were to be transcribed for each primary task. Every participant had to process all speech of all four tasks. Information of the task condition itself (BN vs CTS and NL vs VL) was allowed to be used.

The N-Best protocol also specified several *conditions*. Primary conditions refer to conditions that have to be performed minimally for a valid submission. Contrastive conditions may be performed optionally, additionally to the primary condition.

One condition was processing time. The primary condition was "unlimited time" for processing, with the only constraint that results had to be submitted within the evaluation period of about one month. Two contrastive processing time conditions were suggested, namely 10× and 1× Real Time. Another condition was training material. The primary condition was a specified subset of the CGN corpus for acoustic training, see Table 1, and a collection of newspaper texts obtained from Mediargus (for Flemish) and PCM (for Dutch) for language model training, see Table 2. A contrastive condition for training was: any material with a creation date before 1 Jan 2007. This condition was made to make sure none of the evaluation data, to be recorded after that date, would be used for training.

Each site had to submit ASR results for one *primary system*, trained in the N-Best primary training condition and run on all primary tasks. Further, any number of contrastive systems were

¹<http://taalunieversum.org/taal/technologie/stevin/documenten/stevin/english>

Table 2: Language model training material in the primary training condition. Source indicates the copyright holder, Partner is the N-Best project partner involved in obtaining the license.

Accent	Source	Partner	Size
NL	PCM	UTwente	360 M
VL	Mediargus	KULeuven	1436 M

Table 3: The sources for the BN tasks in N-Best 2008, and the total duration per source. ‘IDs’ are the numerical identity used to make the show anonymous.

Source	Show	IDs	dur. (min)
BNR	Nieuwsradio	1, 8, 9, 14	28.7
NOS	Radio1 Journaal	2, 13	19.8
NOS	8-uur Journaal	3, 7, 12, 15	54.4
NOS	1-uur Journaal	11	8.9
NOS	Buitenhof	4, 5	10.5
NOS	NOVA	6, 10	11.5
VRT	Koppen	1, 4	16.2
VRT	De zevende dag	2	8.5
VRT	Terzake	3, 11	21.9
VRT	De Ochtend	5, 7, 13	23.6
VRT	Villa Politica	6	4.0
VRT	Vandaag	8, 10, 14	17.2
VRT	Journaal	9, 12	31.4

allowed to be submitted for any subset of the primary tasks, obtained using any contrastive condition, e.g., contrastive technology, training (but respecting the LM source date), or processing time constraint.

The evaluation plan [4] specifies how the hypothesis is scored, some worth mentioning here are:

non-lexical events such as hesitations, filled pauses, coughs, are not scored, but can still lead to insertions in the hypothesis transcription

numbers are to be written out in full, according to specified rules of grouping into single words.

compound words should not be split

capitalization of proper nouns should be correct, e.g., “de tweede kamer links” (the second room on the left) vs. “een debat in de Tweede Kamer” (a debate in the parliament).

accented words should be written as such if leaving out accents lead to ambiguities, e.g., “een blauwe loge” (a blue lodge) vs. “niet één logé” (not a single lodger)

initials and titles are separate words, e.g., “prof. dr. ir. P. Akkermans.”

3. Evaluation test material

3.1. Broadcast News

Material for the BN tasks of the evaluation were obtained directly from the copyright holders, and license agreements were set up so that this material could be used for this evaluation and further be distributed for research purposes by the Dutch Language Union. The radio and television sources for NL and VL accents are shown in Table 3.

Table 4: Gender balance of the CTS speakers.

Accent	Male	Female
NL	11	10
VL	8	14

Table 5: Total duration and number of words in the evaluation for the four primary tasks.

	NL		VL	
	dur	N_w	dur	N_w
BN	2.23h	24435	2.14h	22496
CTS	2.80h	17746	2.54h	16276

3.2. Conversational Telephone Speech

Material for CTS was recorded under auspices of SPEX. Subjects were recruited from a wide range of regions in Flanders and the Netherlands in accordance with the CGN design [3]. Recruitment strategies made it possible that conversation partners were familiar with each other. Subjects were given a choice of topics to discuss from a predefined list of topics, inspired by the Switchboard data collection [5], but they were allowed to discuss other topics if they would wish so. Some subjects participated in more than one conversation, but for the evaluation a maximum of one conversation per subject was used. In recruitment it was difficult to maintain gender balance, specifically for Flemish. Gender statistics of the speakers are given in Table 4.

3.3. Reference transcriptions

In total about 3 hours of material per task condition were recorded. Transcriptions of all BN and CTS recordings were made by SPEX, according to a protocol very similar to the one used in producing CGN. TNO then selected segments from this data, using criteria such as speaker sex and regional distribution for CTS, and de-selecting advertisements, speech in other languages than Dutch, and linguistically less challenging parts of the show such as weather forecast from the BN material. This amounted to about 2 hours per task condition. This material was sent back to SPEX for a verification round of the transcription. Each verification was carried out by another transcriber than the one who made the original transcription. Finally, TNO listened to all evaluation data prior to sending evaluation material out to the ASR participants, fixing some last segmentation errors. Total duration of speech is indicated in Table 5

4. Performance measure and Scoring

The primary evaluation measure is the Word Error Rate (WER, fraction of words either substituted, deleted or inserted w.r.t. the reference transcription), as calculated by the NIST `sclite` scoring software. This effectively aligns the hypothesized transcription with the reference transcription using a dynamic programming algorithm with weights 3, 3, and 4 for deletions, insertions and substitutions, respectively.

The evaluation audio material was augmented with *Unpartitioned Evaluation Map* files [6] indicating which parts of the audio files needed to be processed. Speech recognition results were to be submitted in NIST CTM format, using UTF-8 encoding. Word alignment was carried out in two steps: first, the tool `asclite` was run in order to perform the basic alignment, using the UEM information to score only relevant segments. Then `asclite` was used to compute the basic performance statistics.

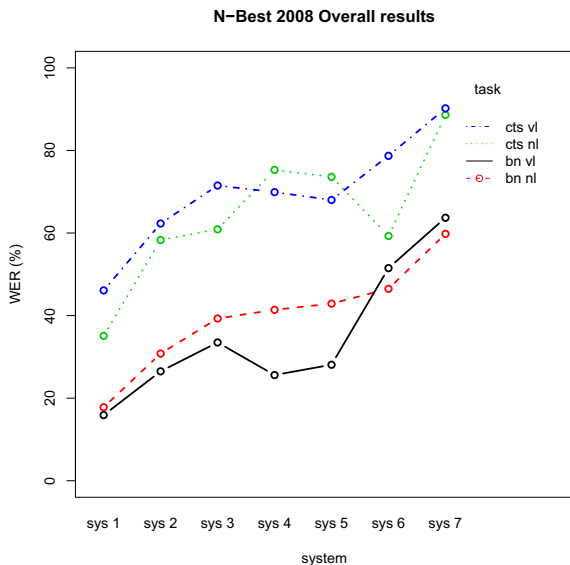


Figure 1: Overall results of N-Best 2008, WER as a function of system and primary task condition. Systems are ordered according to average WER over tasks, lines connecting points are just guides for the eye.

TNO produced initial scoring results one week after the submission deadline, after which a month adjudication period followed. TNO did not only consider adjudications to the scoring suggested by the ASR participants, but also inspected all the word alignments of all the primary system submissions made, specifically for capitalization errors, compounding or even spelling errors in the transcription. It is interesting to remark here, that the best performing system helped finding a significant number of these errors in the reference transcription.

5. Results and Discussion

In accordance with the evaluation protocol, we have made the system names in this publication anonymous. Different sites submitted different number of systems. “Sys 3” ran three different LVCSR systems and four different runs of its main system. “Sys 1” submitted an “unlimited time” contrastive systems for CTS (the primary system happened to be a 10× real time system—which of course also meets the “unlimited time” primary condition requirement). “Sys 2” ran a single-pass system contrasting its multi-pass primary system, and “sys 5” ran a contrastive language model system. In total, 58 systems were submitted distributed over 7 sites and the 4 tasks. In this paper we only report on the primary submission of each site.

Not all sites were able to deliver results before the deadline, specifically, “sys 4” submitted after the reference transcriptions had been made available.

The overall error rates for all tasks and all submitted systems are shown in Table 6 and Fig. 1. We can observe that quite consistently CTS gives higher error rates than BN. This is in line with results obtained in English [7] and are likely due to a higher degree of spontaneity in conversational speech compared to mostly prepared speech style in BN, a possible overlap in speakers in BN between training and test, and a better match for the language model between training and test for BN.

Table 6: Overall results of N-Best 2008. Figures indicate the WER, in %.

	bn nl	bn vl	cts nl	cts vl	Average
sys 1	17.8	15.9	35.1	46.1	28.7
sys 2	30.8	26.5	58.3	62.3	44.5
sys 3	39.3	33.5	60.9	71.5	51.3
sys 4	41.4	25.6	75.3	69.9	53.0
sys 5	42.9	28.1	73.6	68.0	53.1
sys 6	46.5	51.5	59.3	78.7	59.0
sys 7	59.8	63.7	88.6	90.2	75.6

Table 7: WER (in %), as plotted in Fig. 2, but separated for NL (top) and VL (bottom) accent regions. Also indicated is the number of words N_w over which the statistics are calculated (‘k’ means 1000). Separate analysis for male and female speakers has been left out here.

	all	clean	spont	tel	back	degr
sys 1	17.8	11.6	20.2	20.8	14.8	20.9
sys 2	30.8	23.3	33.4	37.0	25.4	32.6
sys 3	39.3	26.2	40.3	62.4	28.5	39.2
sys 4	41.2	25.9	45.8	57.5	33.0	42.5
sys 5	42.9	27.1	49.0	58.0	33.2	41.4
sys 6	46.5	34.8	49.9	61.4	41.9	44.2
sys 7	59.8	51.0	64.8	66.4	53.4	56.3
N_w	24k4	7k2	10k2	3k8	358	2k9
sys 1	15.9	8.5	16.6	12.5	17.5	18.5
sys 2	26.5	16.6	27.6	17.8	28.1	30.4
sys 3	33.5	18.1	35.0	45.9	33.3	35.2
sys 4	25.6	13.6	26.5	27.4	27.2	29.4
sys 5	28.1	16.4	29.5	30.1	29.2	30.1
sys 6	51.5	38.8	52.0	56.8	59.4	54.9
sys 7	63.7	59.1	61.4	57.5	72.2	73.4
N_w	22k5	2k6	13k7	873	869	4k4

Word error rates are generally higher than reported for English (around 10% for BN, 15% for CTS), which may be due to a variety of factors. First, only two ASR sites had extended experience in LVCSR evaluations, which gives a lower fraction of high performance results. Second, this is the first evaluation for Dutch, and the evaluation material (acquired new) was quite different from training and developments test material (both obtained from CGN). However, the word error rates obtained for BN are lower than what we reported earlier [8] for self-conducted evaluations on other Dutch speech material: apparently the competitive nature of the evaluation paradigm has brought the best (out of) researchers.

The differences in performance between sites are quite consistent across task. At the final workshop it turned out that some sites had focused at particular tasks, which may be appreciated from deviations in the general trend in Fig. 1.

5.1. Focus conditions

We have also analyzed the BN results in terms of standard NIST BN “focus conditions”: clean speech, spontaneous speech (labelled ‘spont’), telephone speech (tel), speech with background noise (back), and degraded speech (degr). In Fig. 2 the results per focus condition are plotted, for each site, averaged over accent. Also an analysis per speaker sex is shown. The per-percent scores details are shown in Table 7.

Over a wide range of system performances 10–60%, the “clean” focus gives rise to much lower WER than the other

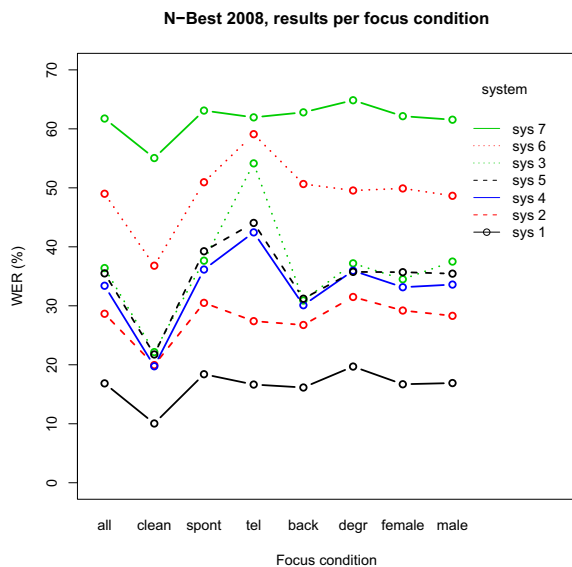


Figure 2: Word error rates for each primary BN submission, analyzed over NIST focus conditions, and separately, sex. For clarity, WERs are averaged for NL and VL accent task conditions.

conditions. Some systems have more problems with the telephone conditions within BN—this may be related to the way CGN training data for BN is organized: for NL these do not contain whole news shows, and telephone data may be missing from these parts.

5.2. Speaker variability

In Fig. 3 we show the range of WER computed per speaker, for the BN tasks and for speakers with more than 500 words. We can observe that the systems with lower WER also show less variability of the WER per speaker. This may perhaps appear to be statistically trivial. Yet, it is interesting to see that systems 2–5 have very similar median performance, while the variance increases. They also show quite different mean speaker performance, cf. Fig. 2, data points “all.” This may suggest that systems 2–5 have progressively less effective speaker-normalisation techniques.

6. Conclusions

The N-Best 2008 evaluation of LVCSR systems is the first evaluation campaign held for the Dutch language. Modelled after NIST-style evaluations, we have defined task and evaluation protocol, collected speech data and produced reference transcriptions, and carried out the evaluation itself. The word error rates of the best system are substantially higher than for well-studied languages such as English, but given that this is the first evaluation for Dutch with evaluation material that is from a different data collection effort than the training material originated from, these results can be considered quite good.

6.1. Acknowledgments

We would like to thank Jon Fiscus and Jerome Ajot from NIST for their very responsive help with the scoring software and other helpful interactions. This work was sponsored by the NTU STEVIN programme.

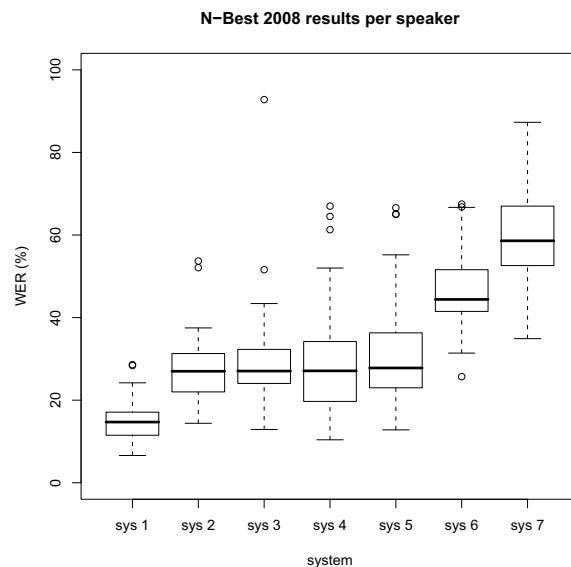


Figure 3: Boxplots of the WER computed per speaker, showing 25 and 75 percentiles (box) and median (line), and range of values (whiskers). The data is aggregated over all BN speakers with more than 500 words in the reference transcription.

7. References

- [1] David Pallett. A look at NIST’s benchmark ASR tests: Past, present, and future. <http://www.nist.gov/speech/history/>, 2003.
- [2] DGA/ELRA. <http://www.afcp-parole.org/ester/docs.html>, 2005.
- [3] N. H. J. Oostdijk and D. Broeder. The spoken dutch corpus and its exploitation environment. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, Budapest, Hungary, 2003.
- [4] David A. van Leeuwen. Evaluation plan for the north- and south-dutch benchmark evaluation of speech recognition technology (n-best 2008). <http://speech.tn.tno.nl/n-best/eval/evalplan.pdf>, 2008.
- [5] J. J. Godfrey, E. C. Holliman, and J. McDaniel. Switchboard: telephone speech corpus for research and development. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 517–520, 1992.
- [6] Jonathan Fiscus. The rich transcription 2006 spring meeting recognition evaluation. <http://www.nist.gov/speech/tests/rt/rt2006/spring/docs/rt06s-meeting-eval-plan-V2.pdf>, 2006.
- [7] Jonathan G. Fiscus, Jerome Ajot, and John S. Garofolo. The rich transcription 2007 meeting recognition evaluation. In *The Joint Proceedings of the CLEAR 2007 and RT 2007 Evaluation Workshops*, volume 4625 of LNCS, pages 373–389. Springer, 2007.
- [8] Judith Kessens and David van Leeuwen. N-best: The Northern and southern dutch Benchmark Evaluation of Speech recognition Technology. In *Proc. Interspeech*, pages 1354–1357, Antwerp, August 2007. ISCA.