Chapter 15
N-Best 2008: A Benchmark Evaluation for Large Vocabulary Speech Recognition in Dutch

David A. van Leeuwen

15.1 Introduction

Automatic Speech Recognition (ASR) is a discipline of engineering that benefits particularly well from formal evaluations. There are several reasons for this. Firstly, speech recognition is basically a pattern recognition task, and to scientifically show that the system works it needs to be tested on fresh material that has never been observed by the system, or indeed the researchers themselves. This means that speech material for testing purposes needs to be collected, which requires quite some effort, but can formally only be used once. It is therefore more efficient if the evaluation material is used to determine the performance of several systems simultaneously, which suggests a common form of this kind of performance benchmarking: that of a formal evaluation. Secondly, after a system evaluation the evaluation material and protocol can be used for future researchers as a benchmark test: algorithms can be developed and tuned to increase performance on the test. By using a well-established formal evaluation protocol performance figures can directly be compared amongst different researchers in the literature, which gives more meaning to the actual figures. Thirdly, a benchmark evaluation gives researchers a clear focus and goal, and appears to stimulate the different research groups to get the best out of their system in a friendly competitive way.

Formal evaluations in speech technology have their origin in the early 1990s of the last century, when the US Advanced Research Projects Agency (ARPA) organised regular evaluations in speech recognition executed by the National Institute of Standards and Technology (NIST) [16], soon followed by speaker [12] and language [13] recognition. In the early years the language of interest for speech recognition invariably was English, but as tasks got harder and performance got

D.A. van Leeuwen (✉)
Centre for Language and Speech Technology, Nijmegen, The Netherlands
e-mail: d.vanleeuwen@let.ru.nl

P. Spyns and J. Odijk (eds.), *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-30910-6_15, © The Author(s) 2013

better, also Arabic and Mandarin became target languages. The NIST evaluation campaigns were so successful that researchers in Europe followed the good example of the US and held their own evaluations of speech technology. One such evaluation was the EU-funded project SQALE[1] [24], in which large vocabulary speech recognition systems (20–65k words) were tested in British and American English, French and German, using read speech. Later, the French *Technolangue* program encompassed *Evalda*, the evaluation of many different human language technologies, among which the ESTER[2] evaluation for Broadcast News speech.

The idea of evaluating technology regularly is the so-called *evaluation paradigm* where system performance is driven to improve over time because researchers compare their approaches in the previous evaluation, gather the best ingredients and implement this in their systems for the next evaluation round. In speech, this paradigm has been implemented most clearly by NIST campaigns, the Technolangue program and Evalita.[3] Other efforts in evaluation, e.g., SQALE and the NFI-TNO Forensic Speaker Recognition Evaluation [23], do no re-occur, and therefore unfortunately do not have the same effect on system performance.

Needless to say, the speech recognition systems require vast amounts of training resources, such as annotated speech material for acoustic models and large quantities of textual material for building language models. These resources were collected and very effectively shared with the research community through the Linguistic Data Consortium (LDC), which again found its European counterpart in the European Language Resources Association (ELRA). In 1998 the Dutch Language Union started a project *Corpus Gesproken Nederlands* (CGN, Spoken Dutch Corpus [14]) aiming at collecting about ten million words of speech as it was spoken by adults in The Netherlands and Flanders at the time. The CGN was created for general linguistic research, and not specifically for the development of a specific speech technology. It thus encompassed many different speech styles, but some of these were indeed suitable for building speech recognition systems for the typical speech recognition task at that time.

Around 2005 there were several research institutions in the low countries that had developed speech recognition systems for the Dutch language [2, 15]. Some were using CGN [20], others used their own databases [10, 15]. The different data used for evaluation and training made it difficult to value the merits of the various systems used. In The Netherlands and Flanders we seemed to be in a situation where there was technology and training material available, but no official speech recognition benchmark evaluation to compare these systems. The STEVIN project *N-Best* aimed at setting up the infrastructure for conducting a benchmark test for large vocabulary ASR in the Dutch language, and collecting data, performing the evaluation and disseminating the results and evaluation data. The acronym N-Best originally is of Dutch origin (*Nederlandse Benchmark Evaluatie voor SpraakTechnologie*) but also

[1] Speech recognition Quality Assessment for Linguistic Engineering

[2] Evaluation des Systèmes de Transcription Enrichie d'émissions Radiophoniques

[3] Evaluation of Natural Language Processing and Speech Tools for Italian, www.evalita.it

has the English interpretation Northern and Southern Dutch Benchmark Evaluation for Speech Technology,[4] expressing the somewhat political wording necessary to indicate the two major language variations in Dutch commonly known as Dutch and Flemish.

This chapter is organised as follows. In Sect. 15.2 the N-Best project is reviewed, then in Sect. 15.3 the evaluation protocol is described. Then, in Sect. 15.4 the evaluation results are presented and discussed.

15.2 The N-Best Project

The project N-Best was funded by the Dutch Language Union research programme STEVIN and consisted of seven partners in three different roles. The coordinator was TNO,[5] responsible of actually carrying out the evaluation. The Nijmegen organisation SPEX[6] was responsible for recording and annotating the evaluation data, and five partners from universities in The Netherlands and Flanders were contributing by developing speech recognition systems for the specific tasks in N-Best and processing the evaluation material. These were ELIS[7] from the Ghent University, ESAT[8] from the Katholieke Universiteit Leuven, the CLST[9] from Radboud University Nijmegen, EWI[10] from the Delft University of Technology, and HMI[11] from the University of Twente.

Despite the competitive nature that a formal evaluation has, N-Best was a collaborative project. In several of the steps that needed to be taken all partners, including the ones with systems under evaluation, collaborated in order to make it feasible for the partners with less experience in evaluation or even large vocabulary speech recognition for Dutch. Most notably, ESAT provided the necessary relation with Mediargus, the supplier for Southern Dutch news paper texts for language model training, and HMI did likewise with their relation with the publisher PCM, the supplier of Northern Dutch newspaper data. Some text-normalising code was shared between partners, and in some cases an entire language model was shared.

[4] Obviously, the term 'Technology' is too broad for a project only dealing with ASR, but this term makes the acronym nicer. Moreover, it can serve as an umbrella name for possible future speech technology evaluations in the low countries.

[5] Netherlands Organisation for Applied Scientific Research

[6] Speech Processing Expertese centre

[7] Electronics and Information Systems department

[8] Department of Electrotechnical Engineering

[9] Centre for Language and Speech Technology

[10] Faculty Electrical Engineering, Mathematics and Computer Science

[11] Human Media Interaction

15.2.1 Specification of the Task and Evaluation Protocol

One of the first things that needed to be established was the definition of the evaluation protocol. Although this was primarily a task of the coordinator, preliminary versions of the document were discussed among all project partners and omissions or errors were pointed out. The result of this process was the publication of the 2008 N-Best evaluation plan [21]. The evaluation plan was inspired by several similar documents from NIST and from the ESTER project, and adapted for the task that was defined for N-Best. The main task was the transcription of Dutch speech, both in the Northern and Southern Dutch variety, and in both the speech styles "Broadcast News" (BN) and "Conversational Telephone Speech" (CTS), amounting to four 'primary tasks'. These styles were well known in the speech recognition community, and well studied in the case of (American) English. Further, the main training condition was to use a specified partition of CGN for acoustical training, and newspaper text provided by partners ESAT and HMI.

15.2.2 Recruitment of Participants

One of the objectives of the N-Best projects was to establish the state-of-the-art of automatic speech recognition for Dutch. In order for this level of performance to be representative of what current technology was capable of, it was important that several of the best laboratory systems take part in the evaluation. Therefore one of the tasks in the N-Best project was to find sites that were willing to participate in N-Best without direct funding from the project. Given the fact that there are not many speakers of Dutch in the world, and that the development of a speech recognition system for a new language requires quite some effort, it was not trivial to find researchers outside the low countries that would participate in the evaluation. Still, we found two teams in Europe that registered: the combination Vecsys[12] Research + Limsi from Paris, France and Brno University of Technology from Brno, Czech Republic. One site registered with the idea of testing a commercial speech recognition system, but had to pull out because the task was too hard.

15.2.3 Testing of the Infrastructure

Because for most ASR partners in the project this was their first formal evaluation, and for TNO it had been over a decade since it had been involved in a speech recognition evaluation, it was decided to have a dry-run in order to test the evaluation

[12]Now Vocopia

process and protocol with respect to file formats, recognition output, file exchange, and scoring. In order to carry this out, some test material was necessary, and for this we utilised parts of the acoustic training material that were marked for development testing. In order to simulate the typical train-test data shift as well as possible within the larger collection of the CGN, the development test data was selected based on recording date. Because the recording date was not available for all parts in the CGN in the standard release from the Dutch *HLT Agency*, a special contract was signed between the coordinator and the HLT Agency, so that the coordinator was able to split off development test material from the training data based on the actual recording date.

Most of the N-Best project partners submitted results for this development test material, and this was scored by the coordinator, such that submission formats and scoring scripts could be tested. The experiences were discussed in an N-Best project workshop. The result was that some of the writing conventions were clarified in the evaluation plan, and that scoring scripts were improved. The development test material, including scoring scripts and the scores of one of the partners, was distributed amongst all N-Best evaluation participants.

15.2.4 Recording, Annotating, and Selection of the Evaluation Data

The evaluation data was recorded by partner SPEX. For the Broadcast News (BN) speech data, material was obtained digitally from the copyright holders, with whom license agreements were set up such that the material could be used for this evaluation, and could further be distributed by the Dutch Language Union. For Conversational Telephone Speech (CTS) data subjects were recruited from a variety of locations within Flanders and The Netherlands. The recruitment strategies allowed for partners in telephone conversations to be familiar with each other – this typically leads to more spontaneous speech which makes it a harder transcription task. In order to stimulate the conversation, subjects were given a topic to discuss from a predefined list of topics, similar to how Switchboard [7] was set up. However, the subjects were free to deviate from this topic. The level of familiarity between subjects and actual topic were not explicitly annotated.

About 3 h of speech for each primary task were recorded. These were all orthographically annotated, using a protocol very similar to the one used in the production of the CGN [8]. This data was sent to the coordinator, who made a further selection in this data, based on criteria such as the speaker's sex and regional variety for CTS, and removing ads and non-Dutch speech from the BN material. After the selection there remained a little over 2 h for each of the four tasks. This selection was then verified by SPEX, with a different transcriber than in the first annotation round. Finally the coordinator listened to all speech prior to sending the data to the participants, and manually remove the last glitches in the data.

15.3 The N-Best Evaluation

The N-Best evaluation was held in April 2008. The evaluation protocol was described in the Evaluation Plan document [21]. The main characteristics of the evaluation protocol are reviewed in this section.

15.3.1 Task

The task in N-Best is that of automatic *transcription* of speech. The speech material is conditioned to one of four domains. The regional variants of Dutch are Northern and Southern Dutch (also known as *Dutch* and *Flemish*). The speech material is obtained from the either radio and television shows (Broadcast News, BN) or telephone conversations (Conversational Telephone Speech, CTS), which are referred to as speech styles. The four primary tasks in N-Best are to automatically transcribe 2 h of speech in each of the four domains formed by the Cartesian product of regional variant and speech style.

15.3.2 Conditions

Several conditions are defined under which the ASR systems should operate. One set of conditions are known as the primary conditions. All participants must submit recognition hypothesis results for each of the primary tasks in the primary condition. Further, sites are encouraged to submit results of any of the task in contrastive operating conditions, where a set of predefined contrastive conditions are suggested. Other important resources for a recognition system, such as pronunciation dictionary, were considered part of the system design and were not controlled or restricted.

15.3.2.1 Primary Conditions

Training material

In the primary condition the training material for acoustic and language models was limited to the material designated and distributed within the N-Best evaluation. The acoustic training material consisted of designated parts of the CGN, as shown in Table 15.1. The language model training material consisted of newspaper text, as distributed by the coordinator. This material was contributed by two of the N-Best partners, also participants, to the evaluation. All language model training material originated from before 1 January 2007, which was the limit for language model training material in *any* of the conditions. The specification of written sources is found in Table 15.2.

Table 15.1 Specification of the acoustic training components of CGN

		Duration (h)	
Speech domain	Component	Northern	Southern
Broadcast news	f: broadcast interviews	42.9	20.9
	i: live commentaries	30.7	12.9
	j: news/reports	8.3	9.1
	k: broadcast news	27.5	8.2
	l: broadcast commentaries	7.9	6.8
	Total	99.4	52.9
Conversational telephone speech	c: switchboard	55.3	36.5
	d: local minidisc	36.7	27.5
	Total	92.0	64.0

Table 15.2 Language modeling training resources for N-Best

Supplier	Newspapers	Years	Size (million words)
PCM	Algemeen Dagblad	2001–2004	66
	Dortsch Dagblad	1999–2000	1.9
	HP de Tijd	1999–2000	0.9
	NRC Handelsblad	1999–2004	82
	Het Parool	1999–2004	57
	Trouw	1999–2004	55
	Vrij Nederland	1999–2000	1.2
	De Volkskrant	1999–2004	94
	Total NL	1999–2004	360
Mediargus	De Morgen	1999–2004	135
	De Standaard	1999–2004	118
	De Tijd	1999–2004	98
	Gazet van Antwerpen	1999–2004	240
	Het Balang van Limburg	1999–2004	106
	Het Laatste Nieuws	1999–2004	284
	Het Nieuwsblad	1999–2004	322
	Het Volk	2000–2004	133
	Total VL	1999–2004	1,436

Processing Time

The primary condition for processing speed was *unlimited time*, with the condition that results needed to be submitted within the deadline, which was 25 days after the data became available. There was no restriction to the number of CPUs or cores that are used to process the data.

15.3.3 Contrastive Conditions

Training Material

Contrastive training conditions could be formed by using any acoustic or language modeling training material, as long as the material originated from before 1 Jan 2007. This is because the evaluation test material was obtained from recordings that were made after this date, and thus we could be reasonably sure that the evaluation material (in speech or text form) did not occur in any training material used.

Processing Time

Contrastive conditions in processing speed could include any speed restriction. In line with other international evaluations, we suggested the specific processing time restrictions of $1\times$ RT (real time) and $10\times$ RT.

15.3.4 Contrastive Systems

Each site was to submit primary task results for at least one system, the primary system. Participants were encouraged to submit results for other, contrastive systems, for any of the tasks in any of the conditions, as long as also primary system results were submitted for these conditions.

15.3.5 Evaluation Measure

The primary evaluation measure of performance was the Word Error Rate (WER), as calculated by NIST `sclite` tools [6].[13] In the determination of the WER non-lexical events (coughs, filled pauses, etc.) were not included in the reference transcription. However, an ASR system would have to indicate these non-lexical events as such if it recognised these events, or these would be counted as insertions.

The evaluation plan [21] specified the way numbers, compound words, acronyms, capitalisation, abbreviation, accents, punctuation should be used in the system's output. Further, relaxed interpretation of spelling was adhered to because of the many spelling reforms the Netherlands and Flanders have experienced in the past.

[13] Available from http://www.itl.nist.gov/iad/mig/tools/

15.3.6 File Formats

The files used in the evaluation were all in standard formats. Audio was distributed in RIFF/WAV files with 8 kHz A-law two-channel encoding for CTS and 16 kHz 16-bit linear PCM encoding for BN. Evaluation control files, specifying which parts of the audio were under evaluation, were in NIST Unpartitioned Evaluation Map (UEM) format [4]. The recognition hypothesis results were expected in UTF-8 encoded CTM files [4]. Specifically, only words in field 7 of type `lex` are considered in computing the WER, other types are ignored.

15.3.7 Evaluation and Adjudication

The speech data were released as a downloadable archive file containing all speech files, together with the corresponding UEM files. Results were due at the coordinator within 25 days. Results arriving late were marked as such. Within a week the coordinator released the first scoring results, including references and alignments, after which there was a 2-week adjudication period. Here, participants could question certain decisions in the scoring. Finally, the coordinator would release the final results.

15.4 Results

15.4.1 Submission Statistics

During the preparations of the evaluation [21], it was decided that in written publications comparative results [22] are to be presented anonymously, but that individual sites can of course present their own results [1,3,9]. This was inspired by the way it goes in the very successful NIST Speaker Recognition campaigns, and the most important reason for N-Best was to make the evaluation more attractive for industrial participants. However, one industrial subscription to the evaluation pulled out at the last moment, so the anonymity in this publication only serves to adhere to original agreements.

There were seven sites participating in the evaluation, including the five ASR sites from the N-Best project. Six of these submitted results before the deadline, totaling 52 submissions distributed over the four primary tasks. Each of the six sites included their primary system in these submissions. One participant (“sys 1”) refrained from receiving the first results until about 3 days after these had been sent to the other five participants, in order to finish two ‘unlimited time’ contrastive runs for their CTS system.

One of the participants (“sys 4”) did not submit results, but refrained from interaction with any of the involved parties, until about 4 months after the official

Table 15.3 Overall results of N-Best 2008. Figures indicate the WER, in %. Systems with * indicate late submissions

	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

deadline, due to unavailability of personnel. This amount of delay is quite unusual in formal evaluations, and it is difficult to guarantee that no information about the evaluation will have reached this participant.

"Sys 3" ran three different ASR systems and four different runs of its main system. "Sys 2" ran a single-pass system contrasting its multi-pass primary system, and "sys 5" ran a contrastive language model system. Finally, 'sys 6' and 'sys 7' only submitted the required minimum of four primary tasks.

15.4.2 Primary Evaluation Results

Results for all seven primary systems in the primary conditions in all four primary tasks are shown in Table 15.3 and are plotted in Fig. 15.1. The systems are numbered in order of the average word error rate for the primary tasks. It should perhaps be noted here that 'sys 1,' showing the lowest word error rates for all tasks, submitted a $10\times$ RT system as primary system results, and had a slightly better performing 'unlimited time' contrastive system, which still is according to the rules.

We can observe from the results that CTS gives higher error rates than BN, which is consistent with results reported for English [5] . Apart from the smaller bandwidth of the audio channel, CTS also contains more spontaneous speech than the more prepared speech style that is characteristic of BN. The acoustics of CTS will also contain more regional variability compared to speech available on radio and television, so therefore the acoustic models have less spectral information to model more widely varying acoustic realisations of the sounds in the language. Another effect that makes BN data have less errors than CTS data is that the majority of the language model training material will match the linguistic content of the BN speech better than that of CTS.

15.4.3 Focus Conditions for Broadcast News Data

NIST has defined standard 'focus conditions' for the various types of speech that may appear in BN material: clean, spontaneous, and telephone speech, speech with

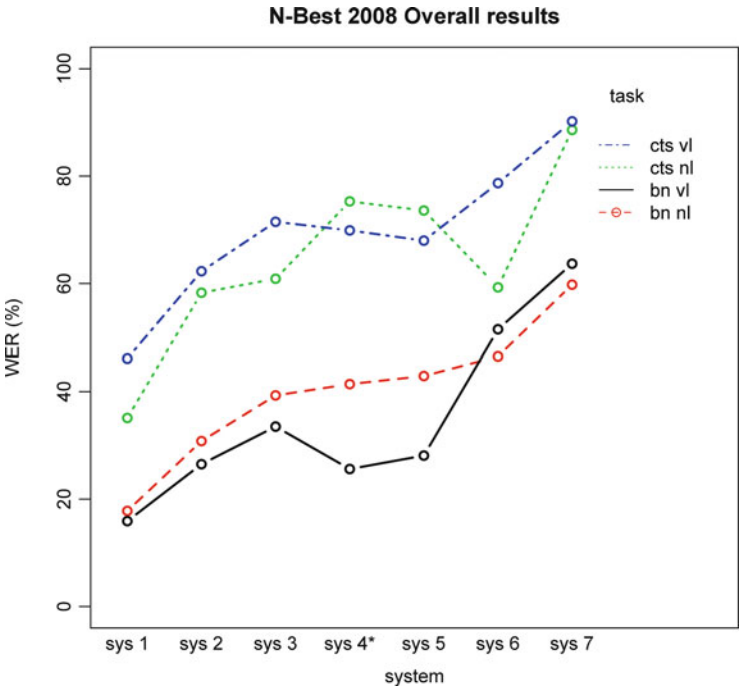

Fig. 15.1 Overall results of N-Best 2008, WER as a function of system and primary task condition. Systems are ordered according to average WER over tasks, lines connecting points are just guides for the eye. Systems with * indicate late submissions

background noise, and degraded speech. SPEX has annotated the test material for these five standard focus conditions, but in the selection criteria for the final evaluation material these conditions were not included. Hence, the amounts of data found in each of the focus conditions is not homogeneously distributed. In Table 15.4 and Fig. 15.2 the WER performance conditioned on focus condition, regional variety and speaker's sex are shown in various combinations.

Even though the performance varies widely over the different systems, ranging 10–60 %, the clean focus condition clearly has lower WER, which is not surprising. Some systems took a particularly big hit with telephone speech in the NL regional variant. This may be resulting from the way the BN training (and therefore, dry run test material) is organised in CGN: contrary to the VL variant, CGN does not contain whole news shows for the NL variant. It is conjectured that the systems that proved particularly vulnerable to telephone speech have been concentrating more on the NL part during development, and may have missed the fact that BN shows may contain this type of speech. This is consistent with the type of errors seen most for these systems in the telephone condition, deletions.

Table 15.4 BN performance expressed in WER (in %), as plotted in Fig. 15.2, but separated for Northern (*left*) and Southern (*right*) regional variants. Also indicated is the number of words N_w over which the statistics are calculated ('k' means 1,000). Systems with * indicate late submissions. Focus conditions are: all, clean speech, spontaneous speech, telephone speech, speech with background noise and degraded speech

NL	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

VL	All	Clean	Spont	Tel	Back	Degr
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

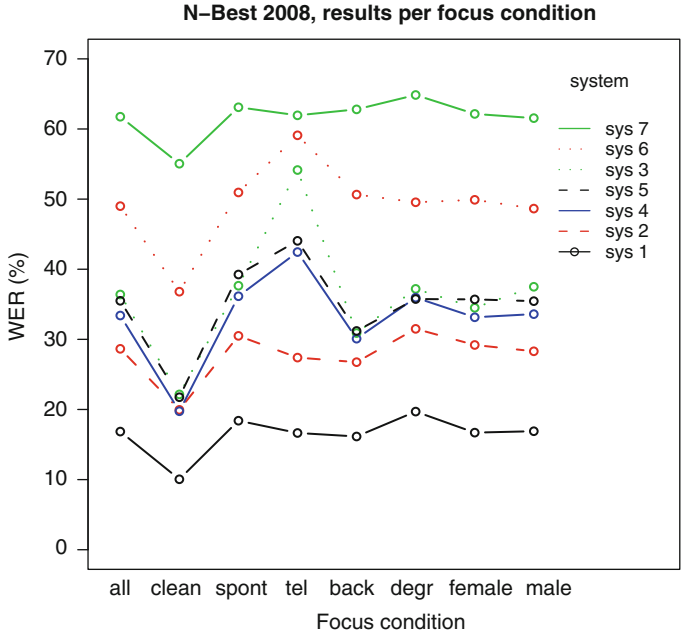

Fig. 15.2 Word error rates for each primary BN submission, analyzed over NIST focus conditions (see Table 15.4 for the legend), and separately, speaker's sex. For clarity, WERs are averaged for NL and VL accent task conditions

However, the performance of telephone speech in BN still is a lot better than in the CTS task for all systems, with notably one exception: that of 'sys 6' for NL. This systems CTS performance is actually better than in the BN telephone focus conditions. This could be explained by 'sys 6' not detecting telephone speech in NL BN data, thus not benefiting from their relatively good CTS NL acoustic models.

15.4.4 Other Accent Effect

Related to this is the analysis of results by origin of the partner. From the N-Best project partners, the partners located in Belgium performed relatively well on Southern Dutch, while the Dutch university performed better on the Northern Dutch variant. This can be appreciated from Fig. 15.3, where the interaction between the participant's home country (North for The Netherlands, South for Belgium) and regional variant of the speech is shown. This is, in a way, similar to the famous 'Other Race Effect' of human face recognition,[14] that is also observed by automatic face recognition systems [18]. We therefore coin this the 'Other Accent Effect.' We have no direct evidence why this is the case, but one reason could be the choice of phone set, the pronunciation dictionary and grapheme-phoneme conversion tools. This is one part of the ASR systems that was not specified as part of the primary training conditions. We can surmise that the researchers had better quality dictionary for their own regional accent than for the other region.

15.4.5 Contrastive System Conditions

Three sites submitted contrasting focus conditions. 'Sys 1' submitted contrasting results showing the effect of processing speed. In Fig. 15.4 it can be seen that faster processing restrictions have a negative effect on performance, but that there probably is hardly any benefit of going beyond 10× RT.

'Sys 2' ran a single-pass system as contrastive to its multi-pass primary system. The results show a quite consistent gain in WER of approximately 10 %-point for all primary tasks when running the multi-pass system (Fig. 15.5).

Finally, 'sys 3' submitted many different contrastive conditions. The main variation was in system architecture, where this site submitted results based on Soft-Sound's 'Abbot' hybrid Neural Net/Hidden Markov Model (HMM) system, [19] the site's own'SHoUT' recogniser [9] and 'Sonic' (University of Colorado, [17]) both pure HMM systems. Using SHoUT, both single and double pass system results were submitted, and additionally a 'bugfix' version of these two were scored by the coordinator. A plot comparing all of these submissions from 'sys 3' is shown in Fig. 15.6. The multi-pass systems did not improve either of the HMM systems very much, about 1 %-point for BN in both accent regions in the case of SHoUT's SMAPLR (structured Maximum A Posteriori Linear Regression) adaptation technique, and about 0.5 %-point for Sonics's CMLLR (Constrained Maximum Likelihood Linear Regression) implementation.

[14] Popularly speaking, the fact that Europeans find it difficult to recognise individual Asians, and vice versa.

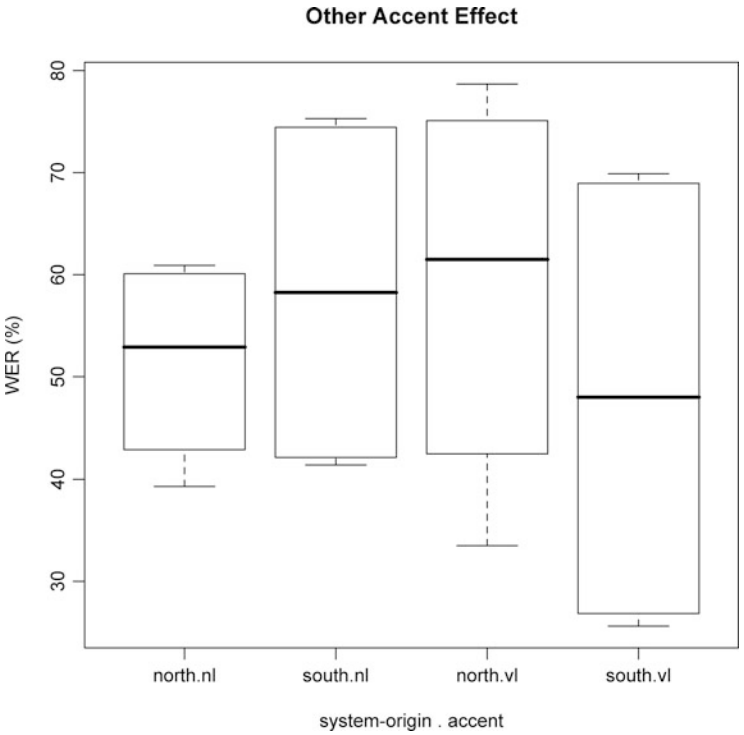

Fig. 15.3 Interaction box plot between the country of origin of the speech recognition system (North/South) and accent of Dutch (Northern – NL/Southern – VL). One system has been left out of the analysis due to extremely high error rates

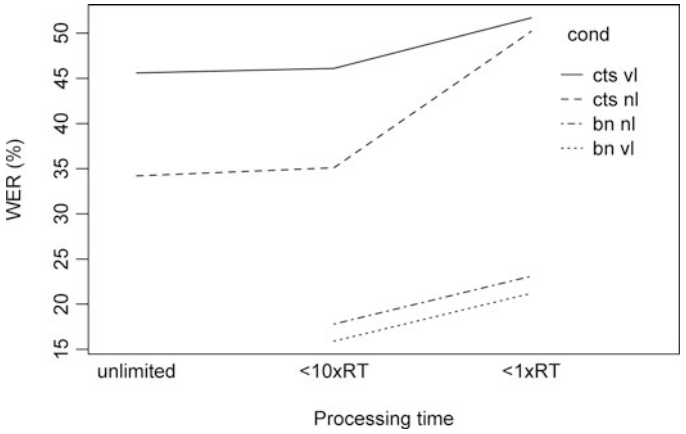

Fig. 15.4 The effect of processing speed restrictions for sys 1

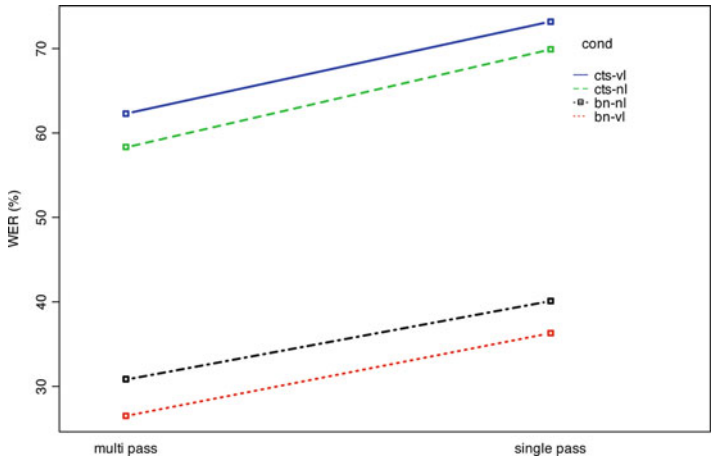

Fig. 15.5 The effect of multiple passes vs. a single pass for sys 2

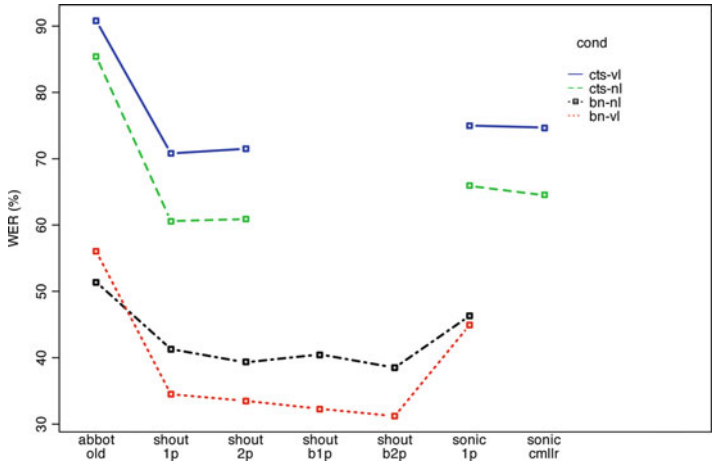

Fig. 15.6 Results for the various submissions of 'sys 3'. The primary system was 'shout 2p'

15.5 Discussion and Conclusions

From the presentations and discussions in the workshop that concluded the evaluation, it became clear that large vocabulary speech recognition in Dutch can be approached quite successfully with technology developed for languages like English, French and German. Dutch is a morphological language with strong compounding (similar to German) and is moderately inflectional. This requires large vocabularies [1, 3] of 300–500k words, but this is not uncommon for languages like German. ESAT reported language models based on morphological analysis [1], which resulted in a moderate reduction in out-of-vocabulary rate and WER during

development for smaller vocabulary sizes. Vecsys Research + Limsi [3] used a common pronunciation dictionary for Northern and Southern Dutch, which was further adapted to the task by including pronunciation frequencies obtained from forced-aligned training data. Further, most sites reported substantial efforts in text-normalisation. For instance, UTwente reported a drop in word error rate from 38.6 to 34.9 % by processing filled pauses, compounds, capitalisation, and numbers [9]. Obviously the rules in the evaluation protocol and the peculiarities of Dutch writing conventions were different enough from other languages to draw considerable efforts from the developers, but did not require radically new approaches to text normalisation. The newspaper text data distributed was from before 2005, while the date limit for contrastive conditions was 1 January 2007. However, no contrastive systems were submitted with more recent language modeling data than 2004.

The results shown in Sect. 15.4 are the outcome of first structural and comparative study of large vocabulary speech recognition for the Dutch language. The main effects observed (difference between BN and CTS, focus condition, number of passes, processing time restrictions) are consistent with what is observed in literature for speech recognition in other languages. The absolute values for the WER of the best performing systems are quite higher than for English, where very low error rates for BN are reported, of the order of magnitude of human transcription errors, and where CTS results have been reported around 15 %. The reason for this probably lies in the size of the training data, which is much smaller within N-Best than for English, where thousands of hours of acoustic training data are available. The fact that nobody submitted a contrastive system with more acoustic training data suggests that this material is not readily available to the researchers. Another reason for the higher error rates for Dutch is the fact that N-Best was the first evaluation, and that Dutch participants were not very experienced in formal evaluations, while the non-Dutch participants were not very experienced in the Dutch language, if at all. From informal inspection of the results we can conclude that the latter factor may be less important than the former.

We would like to note that in inspecting the alignments of hypothesised results with the reference transcriptions, the best performing system ‘sys 1’ caused us to notice several mistakes in the reference transcription where grammatical spelling rules or compound words were involved. We found this quite remarkable. At the same time, the scoring process details and the adjudication issues brought several difficult grammatical construction variants to the surface. Examples are *er aan* vs. *eraan*, *te veel* vs. *teveel* and *ervan uitgaan* vs. *er vanuit gaan*. The different compounding solutions in Dutch are quite hard to choose from, even for a native Dutch scholar. Although we were very lenient in the scoring process towards these issues, and spent a lot of time painstakingly checking every hypothesised error, the effect on the total WER typically was only 0.2 %-point.

Interesting may be the ‘Other Accent Effect’ observed in within the N-Best partners, that the performance for the task in their own regional language variant were better, relatively, than in the other variant. This subtle manifestation of a preference for ones own accent, even through ones own system performance, can

be compared to the 'Other Race Effect' for automatic face recognition fusion algorithms [18].

Concluding, the N-Best project can be said successful in setting up the infrastructure for a benchmark evaluation for Dutch ASR systems. The evaluation data and scoring script can be obtained from the Dutch Language Union through its data distribution agency, the HLT Agency. This includes the scores and recognition hypothesis files of the best scoring system, allowing future researchers to compare the output of their own recognition systems for Dutch to the state-of the art of 2008. The evaluation has generated at least five papers in conference proceedings [1, 3, 9, 11, 22]. It remains to be seen if follow-on evaluations of Dutch speech recognition sparks enough enthusiasm from sponsors and participating developers to be realised. This would change the N-Best evaluation from a once-off benchmark evaluation of the state of the art of Dutch ASR in 2008 to a campaign fitting in the 'evaluation paradigm' with all the benefits of exchange of knowledge and the drive to better performing speech recognition systems.

References

1. Demuynck, K., Puurula, A., Van Compernolle, D., Wambacq, P.: The ESAT 2008 system for N-Best Dutch speech recognition benchmark. In: Proceedings of ASRU, Merano, pp. 339–344 (2009)
2. Demuynck, K., Duchateau, J., Van Compernolle, D., Wambacq, P.: An efficient search space representation for large vocabulary continuous speech recognition. Speech Commun. **30**(1):37–53 (2000)
3. Despres, J., Fousek, P., Gauvain, J.-L., Gay, S., Josse, Y., Lamel, L., Messaoudi, A.: Modeling Northern and Southern varieties of Dutch for STT. In: Proceedings of Interspeech, Brighton, pp. 96–99. ISCA (2009)
4. Fiscus, J.: 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)
5. Fiscus, J.G., Ajot, J., Garofolo, J.S.: The rich transcription 2007 meeting recognition evaluation. In: The Joint Proceedings of the CLEAR 2007 and RT 2007 Evaluation Workshops. Volume 4625 of LNCS, Baltimore, pp. 373–389, Springer (2007)
6. Fiscus, J.G., Ajot, J., Radde, N., Laprun, C.: Multiple dimension levenshtein edit distance calculations for evaluating automatic speech recognition systems during simultaneous speech. In: Proceedings LREC, Genoa, pp. 803–808. ELRA (2006)
7. Godfrey, J.J., Holliman, E.C., McDaniel, J.: Switchboard: telephone speech corpus for research and development. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), San Francisco, pp. 517–520 (1992)
8. Goedertier, W., Goddijn, S., Martens, J.-P.: Orthographic transcription of the Spoken Dutch Corpus. In: Proceedings of the LREC, Athens, pp. 909–914 (2000)
9. Huijbregts, M., Ordelman, R., Werff, L., Jong, F.M.G.: SHoUT, the University of Twente submission to the N-Best 2008 speech recognition evaluation for Dutch. In: Proceedings of Interspeech, Brighton, pp. 2575–2578. ISCA (2009)

10. Huijbregts, M.A.H., Ordelman, R.J.F., de Jong, F.M.G.: A spoken document retrieval application in the oral history domain. In: Proceedings of 10th International Conference Speech and Computer, Patras, pp. 699–702. University of Patras (2005)
11. Kessens, J., van Leeuwen, D.: N-Best: the Northern and southern dutch Benchmark Evaluation of Speech recognition Technology. In: Proceedings Interspeech, pp. 1354–1357, Antwerp, August 2007. ISCA.
12. Martin, A.F., Greenberg, C.S.: The NIST 2010 speaker recognition evaluation. In: Proceedings of Interspeech, Makuhari, pp. 2726–2729. ISCA (2010)
13. Martin, A.F., Le, A.N.: NIST 2007 language recognition evaluation. In: Proceedings of Speaker and Language Odyssey, Stellenbosch, South Afrika. IEEE (2008)
14. Oostdijk, N.H.J., Broeder, D.: The Spoken Dutch Corpus and its exploitation environment. In: Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), Budapest (2003)
15. Ordelman, R.: Dutch speech recognition in multimedia information retrieval. PhD thesis, University of Twente (2003)
16. Pallett, D.: A look at NIST's benchmark ASR tests: Past, present, and future. http://www.nist.gov/speech/history/ (2003)
17. Pellom, B.: Sonic: the university of colorado continuous speech recognizer. Technical Report TR-CSLR-2001-01, University of Colorado, Boulder, March 2001
18. Phillips, P.J., Narvekar, A., Jiang, F., O'Toole, A.J.: An other-race effect for face recognition algorithms. ACM Trans. Appl. Percept. **8**(14), ART14 (2010)
19. Robinson, T., Hochberg, M., Renals, S.: The Use of Recurrent Networks in Continuous Speech Recognition, Chapter 7, pp. 233–258. Kluwer, Boston (1996)
20. Stouten, F., Duchateau, J., Martens, J.-P., Wambacq, P.: Coping with disfluencies in spontaneous speech recognition: acoustic detection and linguistic context manipulation. Speech Commun. **48**, 1590–1606 (2006)
21. van Leeuwen, D.A.: Evaluation plan for the North- and south-dutch Benchmark Evaluation of Speech recognition Technology (N-Best 2008). http://speech.tm.tno.nl/n-best/eval/evalplan.pdf (2008)
22. van Leeuwen, D.A., Kessens, J., Sanders, E., van den Heuvel, H.: Results of the N-Best 2008 Dutch speech recognition evaluation. In: Proceedings of the Interspeech, Brighton, Sept. 2009, pp. 2571–2574. ISCA (2009)
23. van Leeuwen, D.A., Martin, A.F., Przybocki, M.A., Bouten, J.S.: NIST and TNO-NFI evaluations of automatic speaker recognition. Comput. Speech Lang. **20**, 128–158 (2006)
24. Young, S.J., Adda-Dekker, M., Aubert, X., Dugast, C., Gauvain, J.-L., Kershaw, D.J., Lamel, L., van Leeuwen, D.A., Pye, D., Robinson, A.J., Steeneken, H.J.M., Woodland, P.C.: Mutilingual large vocabulary speech recognition: the European SQALE project. Comput. Speech Lang. **11**, 73–89 (1997)