Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 15 (1991), 27-41.

EVALUATING THE PERFORMANCE OF SPEECH TECHNOLOGY SYSTEMS*

Louis C.W. Pols

Abstract

Present knowledge of spoken language (speech) combined with present status of technology already allows for the development of certain hardware and software systems that are capable to store and forward speech, to generate speech from text, to recognize the speaker who said something, and to recognize what has been said. Unfortunately, none of these speech technology systems is functioning perfectly well under all circumstances, so, scientists and system designers still have to work hard to improve the performance of their systems. Similarly, buyers of such products want to know what value they get for their money. Both in the development phase and in the product evaluation phase, good methods are required to measure the performance in a, preferably standardized, diagnostic and comparative way. Since the technology itself is rather new, also the methods to evaluate the performance of these systems are still under development, and constitute a research area in its own merits. We will report on the development of these evaluation methods and on the results achieved so far.

1 System Performance

A speech technology system is generally developed in order to analyze, code, generate, or recognize speech. This is done with a certain goal in mind, such as the recognition of connected digits (e.g. product code) spoken over the telephone by many different speakers of English, or a spoken newspaper for visually handicapped people, or a low bit-rate speech coder for mobile radio use, or a speech interface in an office environment, or perhaps as a basic research tool to study how acoustic-phonetic knowledge can best be incorporated in a stochastically-based recognizer. In all these cases the optimal approach has to be chosen, and the system has to go through several steps for implementation, training, and testing. Each step requires informal evaluation by the designer before he can proceed to the next step. However, at certain points in the design phase, and most definitely at the completion of the system, a formal evaluation is required. This defines the performance in an unambiguous way. It is no longer sufficient then to say 'it sounds better than before', or 'the recognition score seems to be somewhat improved'.

What is then required are well defined and generally agreed figures of merit. This may sound simple but it can become quite a problem. Take, for instance, a speech synthesizer or, more specifically, a system that pretends to generate intelligible and natural-sounding speech, from whatever text presented to it, in a specific language. For such a system there are many different aspects that should be considered, ranging from

* Preliminary version of a paper presented for publication in The Encyclopedia of Language & Linguistics.

text interpretation and grapheme-to-phoneme conversion, via choice of speaking style, rhythm, voice characteristics, and intonation, to intelligibility at word and phoneme level. All these aspects jointly determine:

- the overall intelligibility and naturalness, as well as
- the adequacy for actual applications (announcements, database information, newspaper reading, spoken e-mail or telex messages), and
- the acceptability by the intended user (cooperative, experienced vs. untrained, naive).

The situation is not always as complex as with the above example of a rule synthesizer. But even for a 'simple' word recognizer (specifically trained for one speaker, with only a small vocabulary of words, each word spoken in isolation), performance cannot unambiguously be defined. The percentage of correctly recognized words seems to be the most straightforward figure of merit. However, apart from the errors caused by substitutions (incorrectly recognized words), what to do then with insertions (word recognition without actual word input), deletions (no recognition at all, although a word was spoken), and rejections (not accepting a correctly spoken word)? How does the system react to non-vocabulary words or spurious sounds? Should one use live input only for testing, or would pre-recorded speech material be allowed? What are the environmental conditions (noise, competing speech, reverberation, telephone channel), how much training is allowed? Instead of counting words correct, one could argue that it is more relevant to specify how graceful the system recovers from an error, or what the actual throughput is in terms of time needed to perform a pre-defined task, for instance in comparison with another input modality, such as typing.

One can imagine that moving from the relatively simple example of isolated word recognition to natural speech understanding in a dialog situation, further complicates a performance definition. For instance, should one count words correct, sentences correct, or correct interpretation of meaning?

As said before, apart from performance of the final system, diagnostic assessment of a system under development is also of prime importance. Relevant aspects of the various subtasks then should be defined and their progress measured. This brings us to the scheme as presented in Table 1.

Table 1. Aspects of performance and system development.

why testing?	aspects of system development
to improve performance	development, training
to define performance	relevant system aspects
to compare performance	testing, competition

Although this contribution is not about speech technology itself, but about evaluating the performance of such systems, it is unavoidable to discuss briefly the various components of these systems in the appropriate sections, in order to understand better the variety of evaluation aspects. We will concentrate mainly on evaluation of recognition (Sect. 2) and of synthesis (Sect. 4). The other two topics, speaker verification (Sect. 3) and speech coding (Sect. 5), have much in common with recognition and synthesis evaluation, respectively, and will therefore get less attention here. Some notice will also be given to text and speech databases since they are essential for training and testing synthesizers as well as recognizers. Within the European Information Technology Program ESPRIT, research project 2589 ('Multi-lingual speech input/output assessment, methodology and standardisation', for short

SAM) has been initiated and intends to focus most of the European speech technology evaluation research. The National Institute for Standards and Technology (NIST) in Gaithersburg, Maryland is the American organization most involved in performance evaluation, with a strong interest in progress in the American DARPA Speech Recognition Program.

2 Automatic Speech Recognition Performance

One of the presently most successful speaker-independent 1000-word continuous-speech recognition systems is the SPHINX system developed by K.-F. Lee at Carnegie Mellon University. Similar systems in the American DARPA research community are developed by research laboratories of BBN, MIT/Lincoln Lab, and SRI, most of them based on Hidden Markov models. Besides that, AT&T Bell Labs has developed several competitive systems, whereas IBM was the first laboratory, and still is one of the leading laboratories, in developing large vocabulary speech recognition systems, not just for American English, but by now for other languages as well. Probably the most advanced commercial system is DragonDictate, it uses 'discrete speech' as input. Some European systems are SPICOS and the Olivetti system.

Now, what is the documented performance of these systems, what do the published scores imply, and what is still missing? Table 2 gives some data.

Table 2. Some performance figures of recently developed large-vocabulary speech-recognition systems.

potential	training		testing		word
vocabulary	sentences	speakers	sent.	speak.	accur.
IBM (Bahl et al., 1989) 5k words	2000	10	50	10	89.0%
SPHINX (Lee, 1989) 997 words	4200	105	150	15	96.2%
AT&T Bell Labs (Lee et al., 1990) 991 words	3200	80	150	15	93.3%
DragonDictate (Baker, 1989) 30k words					25-60 words/min. throughput

All these systems are under continuous development and their performance steadily improves, so the scores themselves should be interpreted with some care and are generally not directly comparable. Objective and uniform benchmark test procedures using (subsets of) a pre-recorded speech database were required by DARPA and provide the most direct comparison material. This database is the so-called DARPA 1000-word resource management database, consisting of 2,800 sentences read by 160 subjects (Price et al., 1988). However, as soon as for another system the language or the vocabulary is different, this advantage of being able to use the same database, no longer exists. For that specific DARPA program the common task was considered to be word recognition in sentences (see Table 4 for an example). This implies that neither understanding of the meaning of the sentences nor a detailed acoustic-phonetic specification were required. Word recognition involves symbol string matching between spoken input and recognized output. This is performed by a dynamic programming algorithm. It requires a standard orthographic representation of the reference string, especially of compound words, acronyms, mixed strings of alpha-numerics, digits, and dates. Homophonic substitutions (e.g. their vs. there) are not counted. An extra complication in word string alignments is that any substitution error can also be interpreted as a deletion plus an insertion. However, in most algorithms this combination gets a higher penalty than a straightforward substitution. Unavoidably this leads to an underestimation of real deletion and insertion errors, although this bias only shows up clearly at very low performance levels. It is suggested to use a weighted score in which substitution errors are counted for full and deletion and insertion errors only for half.

The word accuracy score, as given in Table 2, is the percentage correct word score minus the percentage insertions. Just as an example, Table 3 gives all the error and accuracy scores for one specific version of the Lee et al. (1990) system.

Table 3. An example of the various error and accuracy scores for one specific system (Lee et al., 1990).

substitutions (S)	3.7%
deletions (D)	2.3%
insertions (I)	0.7%
total error (E = S + D + I)	6.7%
word accuracy (100 - E = C - I)	93.3%
% correct (C = 100 - S - D, ignoring I)	94.0%
sentence accuracy (all words in the sentence correct, not necessarily equal to semantic accuracy)	66.0%

Table 4 gives an example of various types of error (substitutions, deletions, and insertions) that can occur in recognizing a sentence. This particular example is rather unambiguous, however, if for instance the spoken word 'recognize' is labeled by the system as 'wreck a nice', then should one count this as one substitution error or as three errors (one substitution plus two insertions)? In comparing the performance of two systems using the same data set, one can do better than just comparing error percentages.

Table 4. Example of various types of error (D=deletion, S=substitution, I=insertion, C=correct) that can occur in recognizing a sentence.

INPUT	on	what	day	-	could	dubuque	arrive	in	port
	|	|	|	|	|	|	|	|	|
RECOGN.AS	-	what	would	it	take	dubuque	arrive	in	port
	|	|	|	|	|	|	|	|	|
ERROR TYPE	D	C	S	I	S	C	C	C	C

All those utterances (in)correctly recognized by both systems, do not differentiate the systems. For that, only the number of utterances correctly recognized by one system but not by the other, is interesting. They determine whether any apparent difference in performance of algorithms is actually statistically significant.

One can easily understand that not every N-words vocabulary is equally difficult. This depends, among other things, upon

- the inherent confusability of the N words in the vocabulary;
- the input constraints imposed by the grammar;
- the language model used.

The so-called test-set perplexity P is a measure for the level of uncertainty given by the grammar. If no grammar at all is applied, the perplexity is equal to the number of words in the lexicon, which is 997 for the DARPA database. If a word-pair grammar is used, this perplexity reduces to about 60. This word-pair grammar specifies whether a word sequence is legal or not. Of all possible 997 x 997 = 994,009 word pairs, only 57,878 actually happen to occur in the available set of 900 sentences, and are thus legal. If also the probability of word w_2 following word w_1 is taken into account, the perplexity further reduces to about 20.

For the Dragon system, Table 2 only specifies a very global performance measure: 25 to 60 words per minute throughput for 'well experienced users' in an interactive transcription task of discretely dictated natural language free text. If the evaluation is not run for diagnostic purposes and only serves to define the usefulness of the system, this throughput measure may well be very feasible.

In the preceding paragraphs we have somewhat concentrated on large vocabulary systems (1000 words and more), but also for cheaper and simpler commercial word recognition devices (about 100 words vocabulary, words spoken in isolation or connected) there are many potential applications, such as in the broad area of command and control, whether or not in combination with hands- and/or eyes-busy conditions. A critical evaluation is then the more useful. Table 5 indicates a number of factors that influence performance (Fourcin et al., 1989). These factors thus have to be specified in any test in order for the results to have any relevance.

Table 5. Factors that influence the performance of speech recognizers.

factor	range
speaking style	isolated, connected words; continuous speech; fluent speech
vocabulary	yes/no; digits; alphabet; 1000 words; 30k words; whole language
speaker population	single speaker; multiple accent, dialect, or language groups
enrolment	vocabulary (in)dependent; speaker (in)dependent; speaker adaptive
dialog structure	voice button; menu; simple artificial language; (pseudo) natural language, limited domain
environment	quiet; over telephone; office; factory

Table 6. The type of test, the data used for testing, and the performance standards and measures presently used for recognizer testing.

TYPE	field test off-line test
INPUT DATA	live utterances pre-recorded speech material
CONSISTING OF	general purpose data: digits, alphabet, command words task-specific data diagnostic test material: minimal word pairs
PERFORMANCE STANDARDS	human performance reference system
PERFORMANCE MEASURES	% recognition rate confusion matrices relative information loss human equivalent noise ratio equivalent vocabulary capacity

Table 6 globally specifies the type of test, the data used for testing, as well as the performance measures applied. Most of these tests are condition-specific and require renewed testing if another vocabulary or another condition is chosen. They have little diagnostic value and require substantial numbers of utterances especially if the performance of the systems improves and still a certain confidence level is required.

Performance figures can be very useful as a criterion for systematically evaluating certain recognition characteristics of (laboratory) systems under development. A good example is the search for the best distance metric to measure the similarity between test utterance and reference utterances.

Presently research is on the way to see whether more general evaluation approaches could be chosen. One of them is the so-called recognizer sensitivity test. With a carefully controlled natural-speech database one can, in principle, cover all sources of variability (e.g. speaker, vocabulary, recording and transmission channel). After choosing a limited set of analysis parameters (such as speaking rate, vocal tract area, temporal word congruence, ratio of peak to average energy, ratio of high to low frequency energy, fundamental frequency, and vocabulary difficulty; Peckham et al., 1990) one can then measure the potential range and distribution for these parameters. If a particular recognizer is then tested with the designed database covering the full range, its performance can, in principle, be predicted for the actual set of field conditions without running a field test.

Another related approach is the search for critical phonetic dimensions which characterize speech in a global manner and yet are critical for recognizer performance differences. Also the use of diagnostic databases that can be systematically manipulated by using synthetic speech is considered as an alternative for testing speech recognizers. The Esprit SAM project is active in all these areas of research.

At the same time one tries to standardize and to automate the existing test procedures as much as possible. Examples are the tests performed some 10 years ago by Doddington and Schalk (1981) and more recently by Nusbaum and Pisoni (1987). Also within the SAM workstation several scoring algorithms are incorporated and easy access to large speech databases is developed. This makes it easier to run identical tests at different laboratories and to compare the results (Steeneken et al., 1989). Most recognizers happen to be very sensitive for fine tuning of threshold and volume settings. This questions the relevance of high scores under optimized conditions, since these can frequently not be reproduced under field conditions.

3 Speaker Verification Performance

Since the rather successful speaker verification system developed in the early seventies by Texas Instruments (TI), there has been slow progress in this area. For an overview paper, see Doddington (1985). The general procedure in speaker verification is as follows: A person claims to be 'A' and the speaker verification system checks whether the voice sample that is generated upon request makes it probable that he is actually person 'A' and can be accepted, or else he is 'not-A' and is rejected. This verification decision is based upon the cumulative Euclidean distance between the features of the speaker's reference frames and those of the time-aligned input frames.

Secure access control was considered to be a major application area for speaker verification, whether or not in combination with other person identity checks, such as weight, badge, signature, finger print, or iriscopy. So far this optimism has not come through, although new chances may come with the use of voice over the telephone for banking transactions, or the use of voice I/O (input/ output) in office applications. Encouraging results have recently been achieved for speaker verification over long distance telephone lines (Naik et al., 1989).

For forensic applications (could the voice on this recording be that of suspect 'A') there is more scientific doubt than support (Bolt et al., 1970).

The original TI system, using multiple 4-word verification phrases, was carefully evaluated over many years of 24 hr. per day operational use. The gross rejection rate of approved users was 0.9%, with a casual impostor acceptance rate of 0.7%. There is always a trade-off between these two error measures. Sometimes the square root of the product of these two scores is used as a single performance measure. Late reaction to prompts, number of people in the entrance booth, time of the day, enrolment, experience, and 'goats' vs. 'sheep' (bad vs. normal performing users) are all aspects that influence the final score. Because for these text-dependent systems the text definition is considered to be part of the system definition, comparative evaluation between systems is difficult. For free-text systems it should be easier to define a benchmark database representative of many different conditions.

4 Speech Synthesis Performance

In sect. 1 we have already indicated that there are many different aspects that influence the quality of a synthesis system producing spoken realizations from text (in print, in file, or from concept). These system components and quality aspects range from text complexity and grapheme-to-phoneme conversion, via voice and sound characteristics, to naturalness and intelligibility. Table 7 gives a summary account.

Table 7. Various system components and quality aspects of a text-to-speech rule-synthesizer.

POSSIBLE COMPONENTS
- text complexity, text pre-processing - lexical search, morphological decomposition, grapheme-to-phoneme conversion, semantic analysis, syntactico-prosodic parsing, phrasing (syntactic boundaries), accentuation (sentence accent), speaking rate and rhythm - intonation, duration, syllable boundary, word stress - selection of unit for acoustic realization, spectro-temporal characteristics - voice characteristics - sound synthesizer - system control strategy
QUALITY ASPECTS
- text interpretation, correct focus words, given/new information - acceptable prosody - word intelligibility - phoneme intelligibility - naturalness - general acceptability

Probably the most widely evaluated text-to-speech (TTS) system is MITalk-79 (Allen et al., 1987) for American English, and the commercialized version of that system called DECtalk, with the male voice Paul.

The Modified Rhyme Test (MRT) was used to measure the phoneme intelligibility. The MRT employs six lists of 50 monosyllabic meaningful words each. Subjects had to choose their response (forced choice) from six possible alternatives such as 'peas, peak, peal, peace, peach, and peat'. The consonant scores (averaged over initial and final consonants) are given in Table 8.

Table 8. Various test results at phoneme, word, and text level, for two speech synthesizers and for natural speech.

type of test	MITalk-79	DECtalk 1.8, Paul	natural speech
MRT, consonant correct score			
six alternatives	93.0	96.8	99.5
open response	75.4	87.1	97.2
DRT, consonant correct score			
2 alternatives		88.8	95.6
idem with SNR=0 db(A)		62.2	79.8
word recognition rate			
Harvard sentences	93.3	95.3	99.2
Haskins sentences	78.7	86.8	97.7
text comprehension test			silent reading
composite correct score	70.3		77.2

Not just the overall correct score, but also the percentage error per consonant class (e.g. stops or nasals), as well as the full confusion matrices, give useful diagnostic information. Later on the same word lists were used in an open response task. The resulting substantially lower recognition scores are also given in Table 8. Another phoneme intelligibility test is the Diagnostic Rhyme Test (DRT), this is also a forced-choice test with just two alternatives per word. Stimuli are words like 'dune' vs. 'tune' for the voicing opposition, or 'knock' vs. 'dock' for the nasality opposition. From the results in Table 8 it is clear that for certain tests there is a danger for a ceiling effect. It is also clear that, despite the good performance of these systems, the human speaker by far outperforms the synthesizer.

The word intelligibility was measured by using two sets of sentences: the 100 so-called Harvard sentences (regular sentences with five keywords each, e.g. 'A pot of tea helps to pass the evening') and the 100 Haskins syntactically correct but semantically anomalous sentences with four keywords (e.g. 'The sick seat grew the chain'); see again Table 8 for the scores.

Also continuous synthetic speech understanding tests were run by using text passages that are commonly used in reading comprehension tests. From the answers to the multiple choice questions, composite scores were derived, both for the listening task as well as for a silent reading task. The average scores in Table 8 show some advantage for reading over listening to synthetic speech, although this advantage disappeared in the second half of the test.

Also word processing and memory load experiments were performed with synthetic speech, such as lexical decision (classifying a stimulus, such as 'parents' or 'peemers', as fast as possible as either a word or a non-word), word recall, and word gating. For an overview, see Pisoni et al. (1985). Apart from native listeners, also non-native listeners and children were used as subjects.

No systematic evaluation of above mentioned systems took place at the text pre-processing (e.g. correct interpretation of punctuation marks, abbreviations, number sequences) and at the linguistic level (e.g. phoneme representation, sentence accent, word stress). Actually, very few tests are available at this level.

Next we will discuss again the various aspects of synthesis evaluation, but this time without much emphasis on the performance of specific systems, as we did above, but with more emphasis on the methodology as such. Table 9 gives an overview of the various distinctions that can be made in evaluating a TTS system.

In terms of the terminology introduced in Table 9, the above-mentioned modified rhyme test is a subjective, diagnostic, laboratory test, operating at the acoustic-phonetic, segmental level. Below, most topics mentioned in Table 9 will subsequently be discussed in some more detail.

4.1 Text pre-processing, grapheme-to-phoneme conversion

Whether the characters (graphemes) in the input text are correctly interpreted can quickly be checked on paper. Abbreviations, punctuation marks, and number sequences generally cause most problems. For one application the details of the text are more important than for another, for instance proofreading versus reading a newspaper.

The performance of the grapheme-to-phoneme conversion component can partly be checked against a large phonematized dictionary of that specific language, although for between-word interaction only a phonematized test database with stress markers will do. For realistic telecommunication applications the correct pronunciation of names, especially those with a foreign origin, is a challenging problem.

Table 9. Distinction that are relevant to be made in evaluating TTS systems.

acoustic-phonetic (signal aspects)	vs.	linguistic (text pre-processing, grapheme-to-phoneme conv., stress assignment)
segmental (phoneme intell.)	vs.	supra-segmental (word and sentence aspects, including prosody)
diagnostic (detailed)	vs.	global (overall measures, e.g. magnitude estimation)
laboratory tests (controlled conditions, usually trained and paid subjects)	vs.	field tests (application-specific, dialog, naive users, adverse conditions)
objective (via physical means)	vs.	subjective (via listener judgments)

4.2 Segmental evaluation

As can be seen in Table 8, the phoneme intelligibility of a good synthesizer is still far below that of natural speech, especially in open response tasks. Therefore, it is most valuable to improve that intelligibility and to define sensitive tests to systematically evaluate that progress. The above-mentioned rhyme tests (MRT and DRT) are not considered to be the best choices, for several reasons. Although the task is easy to perform and requires little training, it is not a very realistic task. Furthermore, once the systems get better there is the danger for ceiling effects. Finally, the rhyme word alternatives have to be defined and agreed upon, which may be posssible for one language but certainly not over languages.

It is better to use nonsense words exclusively, or in combination with meaningful words, of just a few fixed forms, with a controlled frequency of occurrence of all phonemes, and an open response task. Phonotactic constraints in the language should be respected. It is suggested in the SAM project to use a basic set consisting of one or more versions of three lists of CV, VCV, and VC words, respectively. The VC-list can be skipped for those languages, such as Italian, for which closed syllables are virtually non-existent. All single consonants (C) in a language in initial, medial, or final position are combined with the three vowels (V) /i, u, a/. So, for instance for Dutch, with 17 initial consonants, the CV list would consist of 51 words.

Additional word lists can be added upon demand, such as C^nVC words to tests the intelligibility of initial consonant clusters (C^n). One can also deliberately combine nonsense words with meaningful words in the list, as well as single consonants with consonant clusters. Once the test method and the test material is well defined, it is possible to use computer-controlled procedures to generate the word lists in appropriately randomized form, to control the presentation of the stimuli, to score the responses (preferably by using input via a computer keyboard), as well as to process the responses (semi-) automatically, and to present the data in graphical and statistical forms.

4.3 Word and sentence level

Because of the predictability of syntactically correct and meaningful sentences, such as the earlier mentioned Harvard sentences, these are not very appropriate as a critical test. That is why the semantically anomalous Haskins sentences were developed. The only drawback of those sentences is that one fixed grammatical structure (the ADJ NOUN1 VERB the NOUN2) is applied exclusively for a fixed set of 100 sentences. Within the Esprit-SAM project a useful multi-lingual extension was worked out, consisting of five different structures that are identical, or at least similar, over languages. With the five forms fixed and a reservoir of words per word category, the number of sentences is virtually unlimited. Table 10 gives two examples for each of the five grammatical structures for several European languages. For more details, see Grice (1989).

Sentences can also be used in a sentence verification task (subjects should decide as fast as possible whether a sentence is a true statement or not), or in a speech reception threshold task with masking noise. In the next paragraph words and sentences will be the carriers of prosody.

Table 10. Example sentences of five different grammatical structures for several European languages (EN=English, SW=Swedish, DU=Dutch, IT=Italian, GE=German, FR=French).

I.	Declarative with adverbial			
EN:	The table	walked	through the blue truth	
SW:	En stol	dog	till ett tomt hus	
II.	Declarative (Haskins structure)			
DU:	Een warm bot	drinkt	de dag	
IT:	La forte via	beve	il giorno	
III.	Imperative			
GE:	Dränge stets	das Garn und den Fuss		
FR:	Tourne peu	la date ou la main		
IV.	Wh question			
FR:	Quand	le text pose-t-il	la fille crue?	
EN:	How	does the day love	the bright word?	
V.	Declarative with relative clause			
IT:	Il piatto	apre	il pesce	che ride
DU:	De vloer	sloot	de vis	die liep

4.4 Prosodic evaluation

Even if phoneme and word intelligibility have reached an acceptable level of performance, synthetic speech may still sound very unnatural, because of lack of natural prosody. This requires the correct assignment and realization of word stress (primary stress, secondary stress, or unstressed, as for instance in 'conduct vs. con'duct), sentence accent (content words that are in focus generally get sentence accent, the other content and function words not), prosodic phrasing (syntactic boundaries), segmental duration ('recognize speech' vs. 'wreck a nice beach'), and intonation contours. Voice characteristics (such as harsh male, sharp female), speaking style (informative, demanding, gentle), rhythm, and speaking rate are still other aspects of importance. The above mentioned correct assignment and realization does not imply that there is just one best prosody, more probably there are a number of acceptable realizations.

Systematic and standardized prosodic evaluation tests do not yet exist, which is not surprising given the complexity of the problem. Methods so far applied are paired-comparison preference judgments, magnitude estimation, and adequacy, naturalness, or appropriateness scores about specific aspects, such as final rise in F_0, or segmental duration.

Within and outside the Esprit-SAM project attempts are being made to separate out various prosodic aspects, for instance, by separating the existence and well-formedness of F_0 contours from their functionality at morphological, syntactic, and discourse level.

It is common practice to test the prosodic rules by using synthetic speech as generated by the system under development. However, the poor segmental quality of that speech material most of the time seriously interferes with the prosodic characteristics under test. The newly developed PSOLA technique (Charpentier & Moulines, 1989) offers very interesting opportunities to impose prosodic rules directly on natural speech with very little loss of quality. With this technique, changing the fundamental frequency and/or the duration locally, becomes a relatively easy thing to do.

4.5 Paragraph level

Although the paragraph level is the kind of test material in which the effects of all TTS-modules are integrated, so that evaluation at this level would constitute the ultimate test of the adequacy of the total TTS-system, studies at this level are scarce. Above, we mentioned already the use of multiple choice questions upon listening to synthesized reading comprehension test material. In another test, short synthesized newspaper articles, of between 30 and 103 words, were presented to visually handicapped people. Text comprehension questions were asked, as well as mean opinion scores on nine bipolar 10-point scales, related to general quality, precision of articulation, accuracy of pronunciation, voice, stress, tempo, liveliness, fluency, and naturalness.

4.6 Overall quality

Next to diagnostic tests at various specific levels, as described above, there is also a need for just ordering two or more systems or algorithms on a scale by asking judgments to listeners. For instance in telecommunication, one uses for that the Mean Opinion Score (MOS), a 5-point scale. More diversity and precision can be reached by using one or more scales with more than 5 points along the scale. For the global evaluation of the speech quality of rule synthesizers, two univariate scaling methods are recommended, either magnitude or categorical estimation. In the magnitude estimation procedure subjects directly estimate one or more aspects (such as acceptability, intelligibility, or naturalness) by assigning a positive number of their own choice to each utterance produced by each system. In categorical estimation a pre-defined n-points scale is used, with or without a reference point. Here each subject should, preferably, just judge one utterance.

4.7 Field tests

One of the few, and recently performed, controlled field tests is carried out in Sweden to assess reactions towards the application of a TTS-system in a spoken daily newspaper for the visually handicapped. The following aspects were taken into account:

- general pattern of the use of control commands (for listening to parts of the text, or for moving around in the text);
- reading time;
- amount of read text;
- reading speed in terms of words per minute;
- immediate and long-term memory for synthesized texts and for the same text presented by human voice;
- amount of experience.

Studies are under way to assess the suitability of synthetic speech for weather forecast telephone service, for telephone directory service, and the like.

4.8 Objective evaluation

Although apparently the human listener is the ultimate judge about synthetic speech quality, it is certainly most time consuming and costly to run again and again subjective evaluation tests. Just as in telecommunication, it would be nice to have instrumental means to measure (aspects of) intelligibility and speech quality. In telecommunication it is mainly the influence of the channel (telephone bandwidth) or the environment (noise, reverberation), and not so much the speech signal itself, that is evaluated. However, with rule-synthesized speech not just global speech characteristics, such as overall signal-to-noise ratio, average F_0, or pause distribution, might be important but also every single spectro-temporal detail. For certain global speech characteristics, comparisons between natural and synthetic speech have already been made. This concerns F_0- and pause-distribution, as well as average bandfilter spectrum. Detailed, frame-by-frame comparison between synthetic speech realization and several natural speech realizations, using pattern-recognition techniques, may be indispensable. It is easy to recognize here the parallel with performance assessment of automatic speech recognition, where the search for critical phonetic dimensions was mentioned (see sect. 2.).

5 Speech Coding Performance

Coded speech is a special kind of synthesized speech. It does not originate from texts such as in TTS-systems, but it is a coded version of utterances that have once been spoken by a human being. The coding is generally required because of communication channel bandwidth demands, or for security reasons (it allows ciphering), or because it allows for digital storage and retrieval. Coding and bitrate reduction is performed in the frequency and/or in the time domain, by applying such techniques as adaptive or differential pulse code modulation, delta modulation, cell packetization, predictive coding, sub-band coding, voice-excited-, formant-, or channel-vocoding. Apart from low bitrate coding, most speech coding does not make use of the fact that the coded acoustic signal is speech. It could just as well have been music or noise. Since coded speech is a more or less accurate copy of natural speech, contrary to TTS, many aspects of natural speech are still preserved. So, things like grapheme-to-phoneme conversion, stress assignment, segmental duration, or prosody do not have to be evaluated. There might just be a more or less severe, general or specific, deterioration in speech quality, resulting in a reduced intelligibility, speaker recognizability, and naturalness.

Coder performance can be measured by subjective and objective testing. All the *subjective* intelligibility tests that were mentioned above with respect to speech synthesizers can of course also be used for speech coders, although here the listening conditions (noise, telephone, mobile radio) generally get more attention. Another important aspect is the recognizability of the speaker, this is not a great issue in most waveform coders, but most vocoded speech looses many speaker characteristics. Overall quality judgments of coded speech are also used, for instance as indicated on a naturalness or quality scale.

Much more so than for TTS-systems, speech coders, including the condition in which they are used, are *objectively* evaluated. Somehow the distortion (in waveform, spectrum, or spectral envelope) is measured between original signal and coded signal. One must realize that in speech coding the original speech signal generally is available at the transmitting side for comparison.

Especially for waveform coders, (segmental) signal-to-noise measures can be used. Other measures concentrate on the short term spectrum or on the LPC-derived spectral envelope. The test signal can either be natural speech or a well-defined signal produced by an artificial source. The amount of modulation preserved in a number of frequency bands is the basis for the so-called speech transmission index (STI). This objective index appears to be highly correlated with the subjective intelligibility for many different (non)linear distortions (Steeneken & Houtgast, 1980)

6 Future Work

So far, in this and other documents, the evaluation of the performance of the various components of speech technology has been treated separately. However, in man-machine communication in the future, several of these components should be integrated in a natural dialog. This full dialog should then be evaluated. Task performance and efficiency will then have to play a much bigger role. Ergonomical aspects should be considered. For instance, the question whether listeners can easily switch from natural speech, to coded speech, and to rule-synthesized speech, has been barely touched. Also the integration of speech and non-speech devices for input/output will become more important. The better the speech technology components in man-machine communication will perform, the closer we will get to a natural dialog.

Bibliography

Allen, J., Hunnicutt, M.S. & Klatt, D.H. (1987): *From text to speech: The MITalk system*, Cambridge University Press, 216 pag.

Bahl, L.R. et al. (14 names altogether) (1989): "Large vocabulary natural language continuous speech recognition", *Proc. IEEE-ICASSP'89*: 465-467.

Baker, J.M. (1989): "DragonDictate^{TM}-30K: Natural language speech recognition with 30.000 words", *Proc. Eurospeech'89*, Vol. **2**: 161-163.

Bolt, R.H. et al. (1970): "Speaker identification by speech spectrograms: A scientist's view of its reliability for legal purposes", *J. Acoust. Soc. Am.* **47**: 597-612.

Charpentier, F. & Moulines, E. (1989): "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Proc. Eurospeech'89*, Vol. **2**: 13-19.

Doddington, G.R. (1985): "Speaker recognition - Identifying people by their voice", *Proc. IEEE* **73**: 1651-1664.

Doddington, G.R. & Schalk, T.B. (1981): "Speech recognition: turning theory to practice", *IEEE Spectrum* **18**: 26-32.

Fourcin, A.J., Harland, G., Barry, W. & Hazan, V. (Eds.) (1989): *Speech input and output assessment. Multilingual methods and standard*", Ellis Horwood Limited, Chichester, 290 pag.

Grice, M. (1989): "Syntactic structures and lexicon requirements for semantically unpredictable sentences in a number of languages", *Proc. ESCA Workshop on 'Speech input/output assessment and speech databases'*, Noordwijkerhout: 1.5.1-1.5.4.

Hunt, M.J. (1990): "Figures of merit for assessing connected-word recognisers", *Speech Comm.* **9(4)**: 329-336.

Lee, C.H., Rabiner, L.R., Pieraccini, R. & Wilpon, J.G. (1990): "Acoustic modeling for large vocabulary speech recognition", *Computer Speech and Language* **4**: 127-165.

Lee, K.-F. (1989): *Automatic speech recognition. The development of the SPHINX system*, Kluwer Academic Publishers, 207 pag.

Lee, K.-F. (1989): "Hidden Markov models: Past, present, and future", *Proc. Eurospeech'89*, Vol. **1**: 148-155.

Naik, J.M., Netsch, L.P. & Doddington, G.R. (1989): "Speaker verification over long distance telephone lines", *Proc. IEEE-ICASSP'89*: 524-527.

Nusbaum, H.C. & Pisoni, D.B. (1987): "Automatic measurement of speech recognition performance: a comparison of six speaker-dependent recognition devices", *Computer Speech and Language* **2**: 87-108.

Pallett, D.S. (1985): "Performance assessment of automatic speech recognizers", *J. Res. Nat'l Bureau of Standards* **90(5)**: 371-387.

Pallett, D.S. (1989): "Benchmark tests for DARPA resource management database performance evaluations", *Proc. IEEE-ICASSP'89*: 536-539.

Peckham, J., Thomas, T. & Frangoulis, E. (1990): "Recogniser sensitivity analysis: A method for assessing the performance of speech recognisers", *Speech Communication* **9(4)**: 317-327.

Pisoni, D.B., Nusbaum, H.C. & Greene, B.G. (1985): "Perception of synthetic speech generated by rule", *Proc. IEEE* **73**: 1665-1676.

Pols, L.C.W. (1990): (Ed.) Special issue on speech input/output assessment and speech databases, *Speech Communication* **9(4)**: 263-388.

Pols. L.C.W. (1992): "Quality assessment of text-to-speech synthesis-by-rule", In S. Furui & M.M. Sondhi (Eds.), *Advances in Speech Signal Processing*, Marcel Dekker Inc., Ch. 13: 387-416.

Pols, L., Bezooijen, R. van, Heugten, B. van, Koopmans-van Beinum, F. & Steeneken, H. (Eds.) (1989): *Proc. ESCA Tutorial Day and Workshop on 'Speech input/output assessment and speech databases'*, 20-23 Sept. 1989, Noordwijkerhout, The Netherlands.

Price, P., Fisher, W.M., Bernstein, J. & Pallett, D.S. (1988): "The DARPA 1000-word resource management database for continuous speech recognition", *Proc. IEEE-ICAASP'88*: 651-654.

Steeneken, H.J.M. & Houtgast, T. (1980): "A physical method for measuring speech-transmission quality", *J. Acoust. Soc. Am.* **67**: 318-326.

Steeneken, H.J.M., Tomlinson, M. & Gauvain, J.L. (1989): "Assessment of two commercial recognizers with the SAM workstation and EUROM-0", *Proc. ESCA Workshop on 'Speech input/output assessment and speech databases'*, Noordwijkerhout: 6.7.1-6.7.4.