Formative Evaluation of Speech Components in Dialog Systems

GSLT speech technology course | fall 2002

Gustav Öquist
Department of Linguistics, Uppsala University
PO Box 527,751 20 Uppsala, Sweden
gustav@stp.ling.uu.se

Abstract

The goal of formative evaluation is to provide information during development about where a given system succeeds or needs improvement, relative to its intended users and use. Spoken dialog systems typically consist of several components that process information about an ongoing dialog. But, there are usually just two components that actually deal with spoken language. The speech recognizer, which receives the speech signal and turns it into input data for the dialog system, and the speech synthesizer, which turns the output data into a speech signal and return it to the user. Since these components make up the verbal interface with which the user interacts, they have a large impact on usability. Formative evaluation of speech components is thus central to the development of spoken dialog systems. This paper will give an introduction to the methodology, as well as measures and metrics, which can be used for formative evaluation of speech components.

1. Introduction

Today it is possible to create, and fruitfully maintain, a spoken dialog between a user and a computer that would have been perceived as science fiction, or at least a good Wizard of Oz simulation, just a few years ago. One could easily say that recent year's progress in computation is what enabled this progress, but what really made the trick has more to do with fruitful collaboration, and not least sharing of results, between researchers from different fields of research than anything else. Evaluation plays a prominent role in the development of dialog systems; one simply needs to know where a system succeeds and where it fails, both relative to its intended users and relative to the different solutions that are, or can be, employed in the system. The goal of formative evaluation is to provide information about where a given system succeeds or needs improvement, relative to its intended users and use, during development.

What a spoken dialog system essentially does is to retrieve a users spoken utterance, make sense out of it, query the appropriate data sources, figure out an adequate reply, and return a spoken utterance that makes sense to the user. In a spoken dialog system there are usually just two components that actually interact with the user. The speech recognizer, which receives the speech signal and turns it into input data for the dialog system, and the speech synthesizer, which turns the output data into a speech signal and return it to the user. Since these components make up the verbal interface with which the users interact, they have a large impact on usability. Formative evaluation of speech components is thus central to the development of spoken dialog systems since it is important to find out how the users feel about the verbal interface well before it is finished.

The paper sets out with an overview of what a typical dialog system does while processing an utterance, which will illustrate how the speech components fit into the overall dialog system design. After that, the methodologies, and measures and metrics, which are pertinent to different kinds of evaluations, are presented. Although this paper is foremost focused on evaluations of speech components that are performed during development, a brief overview of how speech components are evaluated in general follows next. After that, the methods and measures that are most commonly used for formative evaluation of speech recognition and syntheses are presented and discussed. Finally, a few concluding remarks and speculations of future directions in speech component evaluation wrap up the paper.

2. Overview of a Spoken Dialog System

To begin with, the speech signal needs to be retrieved and turned it into a data structure that can be used by the subsequent system. The component that does this is usually called the speech recognizer or Automatic Speech Recognizer (ASR). What it does is to transform the incoming continuous waveform into intermittent units that can be matched against discrete acoustic templates, usually representing phones in the language of use or other acoustic elements of interest (Blomberg and Elenius, 2000). Depending on the success of this process, a more or less accurate representation of a word can be retrieved, which in turn can be aggregated into a more or less accurate representation of an utterance. The aggregated accuracy scores may be good to keep a record of as these can be used to discriminate a word, or a sentence, from another. The output from the speech recognizer is usually a text string or a set of strings, selected on basis of the utterances that gave the best accuracy scores.

The next step is to make sense of the input given from the speech recognizer. The component that does this is usually called the language analyzer. What it does is to analyse the output from the speech recognition and translate it into a representation that is appropriate for the dialog system to base decisions on, which usually is some form of semantic representation. This process typically involves syntactic parsing that yields constituent structures with associated semantic representations, but it can also use much simpler techniques such as keyword or phrase spotting. The output from the language analyzer is a semantic representation, or a set of alternative semantic representations, that captures the meaning expressed in the user's utterance (McTear 2002).

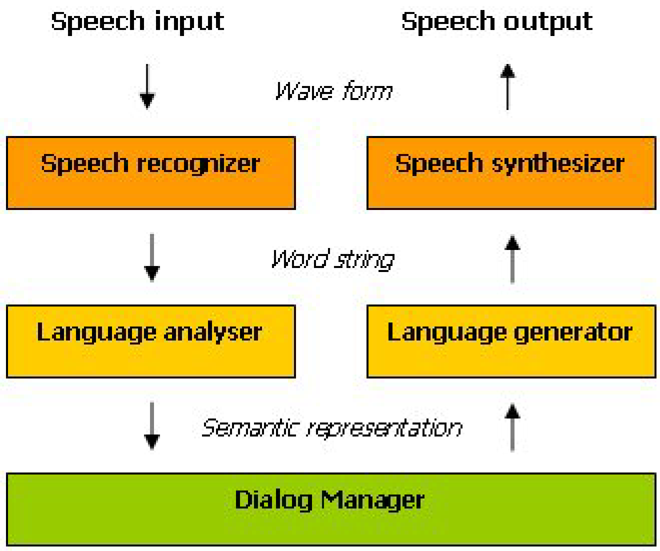

Figure 1. Overview of the components in a spoken dialog system with intermediate representations

At this stage we have come to the heart and soul of the dialog system, the dialog manager. This component both keeps track of the developing dialog as well as promote further dialog by finding answers and formulating questions. The first step in the management process is to interpret the input in the light of previous utterances to see if the system can find an appropriate answer. To find answers the dialog manager needs to communicate with the underlying applications that the system supports, for example a database or an expert system. The second step in the process is to figure out a reply, for example what question to ask next, or what answer to return. The dialogue manager also has to deal with a fair amount of misunderstandings, sometimes caused by the user, sometimes caused by the components of the recognition and understanding process (McTear 2002). The planning of the reply is usually based on models of how a dialog should flow as well as on records of what has been accomplished in the dialog so far. The output from the dialog manager depends on the implementation, but it is usually a semantic representation that conveys the meaning of what the system should reply.

Now the dialog system is ready to formulate an answer that can be returned to the user. The component that does is usually called the language generator. What it essentially does is to turn the semantic representation produced by dialog manager into a string of words, often in plain text. This process consists of mapping intentions in the form of semantic representations to lexical units, which in turned can be assembled into a meaningful sentence according to a grammar (Baptist and Seneff, 2000). In many senses this component can be seen as a reversed version of the language analyzer, which now has to produce meaningful sentences instead of finding them.

The last component in the dialog system transforms the systems answer into a speech signal that can be delivered to the user. This component is usually called the speech synthesizer, or if it operates on plain text, the Text-To-Speech (TTS) component. The first step in the speech synthesis is to generate a representation from which synthesis can take place. Such a representation typically includes information on the actual phonemes to be uttered, which words to emphasize, and how to sensibly chunk up sentences into prosodic phrases (Sprout, 1999). The second step involves synthesizing a waveform given the symbolic representation of the utterance. This is usually done in one of three ways. The least advanced is canned speech, where the output consists of pre-recorded utterances. A more advanced approach is concatenate synthesis, where the speech is produced by pasting together either smaller units, as phones or diphones, or larger units, as individual words. A very advanced way of doing it is called formant synthesis. The speech is then produced by a set of filters and sound sources. The sources can be modelled after real voices. By changing the filter frequencies it is then possible to produce a variety of phones and dynamically produce different voices and voice qualities (Carlson and Granström, 1997).

All dialog systems are component based, not necessarily at the computational level, but certainly at the conceptual. The foremost reason for this is that the system would be too complex to put together otherwise. Each of the components mentioned before requires specific scientific knowledge to construct, and the development of just one of them represents a challenging research field in its own right (McTear, 2002). The components used in a specific dialog system are not necessarily the same as those presented here (Figure 1), neither do they have to be put together in the same way. Some dialog systems use a more integrated design, whereas others some use a more distributed design. However, regardless of approach, all of the utilized components have to be integrated and work efficiently together for a dialog system to function as a whole. The efficiency of the overall system is always what is of most importance, but a system can never be more efficient than its least efficient component, and thus the efficiency of the individual components is important to know of. Evaluation thus plays a crucial role in development of dialog systems and speech components.

3. Evaluation Methodology

The goal with any evaluation is always to answer a set of questions, and it is the questions asked which determine how an evaluation should be performed. In the Survey of the State of the Art in Human Language Technology, Hirschman and Thompson (1997) distinguishes the following three types of evaluations appropriate to different goals:

Adequacy Evaluation — The goal is to find out how well a system, or a component of a system, does what is required to do, relative to the tasks and users at hand. The process of performing an adequacy evaluation requires considerable knowledge of by whom, how and where the system is going to be used.

Diagnostic Evaluation — The goal is to get profile of how well the system, or a component of a system, works relative to a closed set of input variables. In speech components where coverage is important, a common development methodology employs a large test suite of exemplary input.

Performance Evaluation

The goal is to get a measurement of the performance of a system, or a component of a system, in one or more specific areas. One usually defines what one is interested in evaluating (criterion), which specific property of system which reflects the criterion (measure), and how to assess the value returned for a given measure (method).

The goal of *formative* evaluation is to give information about where a given system succeeds or needs improvement relative to its intended use during development, as opposed to *summative* evaluation, which rates systems relative to each other, or some gold standard such as human performance. Formative evaluation is in a wide sense closest related to adequacy evaluation, since the goal is to ensure that the finished system is adequate. However, in practice it takes the form of repeated diagnostic evaluation of subsequent versions of the same system, on a material that is representative for the intended users and tasks. The diagnostic evaluations are in turn based on performance evaluations, which give information about where the system succeeds or needs improvement. Formative evaluation thus combines all three types of evaluations to give information about adequacy by repetitive diagnostic evaluations of performance on material representative of the intended use.

When systems have a number of identifiable components associated with the stages in the processing they perform, as a dialog system evidently has, it is important to be clear of whether it is the system as a whole, or each component independently, that is being evaluated. In the case here, when we are interested in the speech recognizer and the speech synthesizer, a further distinction between intrinsic and extrinsic evaluation must be respected. Do we look at how a particular component works in its own terms, *intrinsic*, or how it contributes to the overall performance of the system, *extrinsic*. When evaluating speech components, intrinsic often relates to performance evaluation whereas extrinsic relates to adequacy evaluation, but these concepts are in no way mutually exclusive. A distinction is also often drawn between evaluations with a glass box and black box evaluation, which differentiates between if it is possible to discriminate which inert variables that affect the result, *glass box*, or if the results only tell about the overall end-to-end performance, *black box* (Hirschman and Thompson 1997). Since formative evaluation should support development it implies a glass box evaluation of the system with both intrinsic and extrinsic evaluation of the components.

The data from an evaluation is usually categorized according to if it is *objective* or *subjective*. The difference basically has to do with a combination of how the data is acquired and what it represents. Objective, or quantitative, data is not supposed to be affected by human judgement. The completion time for a task as it is recorded by a computer is an example of objective data. Subjective, or qualitative, data is on the other hand supposed to be affected by human judgement. The completion time for a task as it is measured by a user is an example of subjective data. Objective data is usually collected by automatic means whereas subjective data is collected by questionnaires and sometimes also interviews. It is usually analysis of combinations of objective and subjective data that gives the most interesting answers, mostly since it can tell about both how a system performed and how the users perceived the performance. However this is not as easy as it sounds, a detailed comparison of algorithms or systems requires a potentially large number of contrasts to be measured. This is at odds with the need to produce statistically significant results, which can only be obtained by using large test sets.

Subjective data collection usually requires quite a lot of subjects in order to get statistically reliable results. This makes subjective data expensive to obtain, but nonetheless is it important to get. The critical point about evaluation of dialog systems, as with software testing in general, is that the longer the intended users are left out, the worse it gets. If user contact during the design phase is avoided, then a large number of problems are likely to emerge when the system eventually go live, and by that time it is very likely to be too difficult and costly to fix. For the development of speech components it is close to impossible to let the users out, since real spoken material and language models are required to build the components, but real care is then instead required in selecting a representative material with regard to the end users, which can be used for both development and evaluation.

4. Evaluation of Speech Components in Dialog Systems

There has been considerable work in evaluation of speech recognizers and speech synthesizers in later years, but in quite different directions, and with quite different results. Evaluation of speech recognition has largely been centred on diagnostic evaluations between several systems on the same test data. Evaluation of speech synthesis has however centred on performance evaluations for individual systems, which is more like formative evaluation, but unfortunately often using different measures and metrics which have made comparisons between different systems hard. The reason for the difference in how the evaluations have been performed may partially be that the methods applied for evaluation of speech recognition differs a lot from those applied to evaluation of speech synthesis, but one other reason may also be there has not been any general strategy as to how evaluation of speech synthesises should be performed. Let us look at these in turn.

One fundamental difference between how speech recognition and speech synthesis can be evaluated has to do with the components representation of input and output (Figure 1). When speech recognition is evaluated, the input representation is a wave form whereas the output is in the form of a symbolic representation. This transformation decreases the information content of the representation. The level of similarity between the symbolic representation and the interpreted wave form is thus quite easy to validate by objective and computational means. On the other hand, for speech synthesis, the input representation is a symbolic representation whereas the output is in the form of a wave form. This transformation increases the information content of the representation. The level of similarity between the interpreted wave form and the symbolic representation is not possible to validate by objective and computational means as we have added information that the quality of only a listener can subjectively assess. The consequence of this is that speech synthesis is much harder to evaluate, but that alone does not account for the lack of common evaluation strategies compared to speech recognition.

One of the main reasons for the development of common evaluation methods for speech recognition, and for the rapid development of the technology as a whole, is funding. In 1984 the U.S. Defense Advanced Research Projects Agency (DARPA) began to finance a large project with the aim to facilitate resources (e.g. dialog system architectures, spoken language corpora, benchmarks, etc.) that could promote development in speech recognition technology. In retrospect one could easily say that the project succeeded, although some also blame it for the narrow focus on technological aspects in speech recognition. Most progress in speech recognition is however a result of DARPA and similar initiatives that has been initiated in Europe and elsewhere, e.g. ESPRIT, SAM, EAGLES, COCOSDA, LDC, NIST (See web resources for links). Most of these programs now also develop resources for the assessment of speech synthesis. Although speech synthesis will remain hard to assess, the situation is likely to improve in the future. There are also a few initiatives to find ways and promote best practice when integrating and evaluating speech components into dialog systems. In Europe there is the DISK working group and in U.S there is the DARPA Communicator project with the PARADISE framework for dialog system evaluation (See web resources for links).

Although evaluation methodologies for the integration of speech components in to the system as a whole, and evaluation of the overall systems are important, this paper will focus on possibilities to formative evaluation of the individual speech components. First we will take a look at speech recognition, and after that we will turn to speech synthesis.

5. Formative Evaluation of Speech Recognition

The aim with speech recognition is to transform a speech signal into a symbolic representation of an utterance. The evaluation thereof is usually based on an assessment of how accurate the output of the speech recognizer is relative to human-produced word transcripts of what the user has actually said. Since the accuracy scores rely on the presence of an accurate reference transcription, it is clearly important to decide what exactly should be captured during transcription (Blomberg and Elenius, 2000). The transcripts must be annotated at the level that will be recognized, such as words or characters, and any artefacts such as noise and music must be marked. Much effort has gone into the development of large speech corpora for system development, training, and evaluation. Some of these corpora are designed for acoustic phonetic research, while others are highly task specific. Nowadays, it is not uncommon to have tens of thousands of sentences available for system training and evaluation (Zue, 1997).

Many of these corpora (e.g., TIMIT, RM, ATIS, and WSJ) were originally collected under the sponsorship of DARPA to spur human language technology development among its contractors, they have nevertheless gained acceptance as standards on which to evaluate speech recognition in English, and stimulated interest for creating similar corpora in other languages as well. The recent availability of a large body of data in the public domain, coupled with the specification of evaluation standards, has resulted in uniform documentation of test results, thus contributing to greater reliability in monitoring progress (Zue, 1997). The most widely used measure of recognition accuracy is the percentage *word error rate* (WER). This measure is usually computed while respecting utterance boundaries, but the errors are usually aggregated across the whole test set to give the overall results. The WER is calculated by aligning the two word strings and counting the total number of words in the test set (N), the number of substitutions (S), deletions (D) and insertions (I). The WER is then given by the following formula:

$$WER = \frac{S + D + I}{N} \; x\,100$$

The standard scoring scheme use a Dynamic Programming (DP) algorithm in which different penalties are applied to substitutions, deletions and insertions. Usually the deletion and insertion penalty is equal and the sum of both larger than the substitution penalty in order to ensure that a substitution is preferred to a deletion plus an insertion. A measure of accuracy that may be used is the Out Of Vocabulary (OOV) measure, which represents the percentage of words that was not recognized although they actually had lexical coverage. In order to retrieve this measure the word forms covered by the lexicon must be known. The WER and OOV values are often combined with confidence scores from the speech recognizer during formative evaluation in order to monitor performance.

In order to keep track of the improvement during development, the selection of the evaluation data should ideally be done in a manner which keeps the level of difficulty constant from test to test. For continuous speech, word perplexity is often used as the primary metric in making this judgement. The word perplexity is a measure of the probability weighted average number of words that may follow after a given word (Blomberg and Elenius, 2000). In order to calculate the perplexity (B), you need to know the entropy (H), which means knowing of the probability of the word sequences in the language of use (W). The perplexity is then given by the following formula:

$$H = -\sum_{\forall W} P(W)^2 \log P(W)$$

$$B = 2^{H}$$

Although the word error rate and the word perplexity are commonly used, they do not tell the whole truth about the accuracy of the speech recognition. They do not say anything about phonological accuracy and neither tell you very much about the level of difficulty of the recognition task. The presence of background noise and channel distortion has a significant impact on recognition accuracy, and must be taken into account as well. Other measures which correlate with the level of difficulty of a recognition task should thus also be used for formative evaluation. Such can be the quality of lexical coverage for the task in question, speech rate, disfluency rate, the amount of mumbling or faint speech, and the rate of use of foreign words. A speech recognizer that is supposed to be a part in dialog system also have some serious constraints on how long time it can process an utterance, recognition completion time is thus important to evaluate as well (Blomberg and Elenius, 2000; Zue et al, 1997).

6. Formative Evaluation of Speech Synthesis

The aim with speech synthesis is to transform a symbolic representation of an utterance into a speech signal. The evaluation thereof is usually based on an assessment of how the output from the speech recognizer sounds. Since it is much harder to enumerate subjective judgments of how a good a sound is, evaluation of speech synthesis is far more difficult than evaluation of speech recognition. Speech quality is generally evaluated in terms of *intelligibility* and *naturalness*. Intelligibility refers to degree to which the speech can be understood. Accuracy of phoneme identification by listeners is usually used as a direct measure of intelligibility, although this neglects the role of prosody in the understanding of the meaning of an utterance. Naturalness is harder to give a straight definition of as it implies an aesthetic assessment. Measuring naturalness usually involves asking listeners to rank speech tokens from the most preferred to the least preferred according to some dimension of naturalness (Sprout, 1999).

Several tests, a few argues too many, for assessing the naturalness and intelligibility of segments, words, and sentences have been developed. They have usually just been used for formative evaluation by the developers and not so often for comparisons between systems. The evaluation methods to test speech quality are usually designed to test speech in general, for example how it sounds over a telephone, but most of them are suitable for synthetic speech as well. It is very difficult, if not impossible; to say which test method that provides the best data. The evaluation procedure is usually done by subjective listening tests (Lemmetty, 1995). A few objective methods exist, but they are mostly concerned with acoustics and the use of them is thus limited. In a speech synthesis system the acoustic characteristics are of course important, but the text pre-processing and linguistic realization also determines the final quality. An extensive survey of methods for speech synthesis evaluation can be found in the DISK guidelines for best practice (Karlsson, 1988). The methods for evaluating intelligibility can be divided into those that only assess the intelligibility of single phonetic segments or those that assess the intelligibility of whole words or sentences. The methods for evaluating naturalness usually assess whole utterances.

Segmental evaluation methods only test the intelligibility single segment or phoneme intelligibility is tested. A very commonly used segmental method is to test the intelligibility of synthetic speech by so called rhyme tests and nonsense words. The rhyme tests have several advantages as the numbers of stimuli is reduced and the test procedure is not time consuming (Lemmetty, 1995). The obtained measure of intelligibility is simply the number of correctly identified words compared to all words and diagnostic information can be given by confusion matrices. Confusion matrices give information how different phonemes are misidentified and help to localize the problem points for development. However, rhyme tests have also some disadvantages. With monosyllabic words only single consonants are tested, the vocabulary is also fixed and public so the system designers may tune their systems for the test, and the listeners might remember the correct answers when participating in the test more than once (Lemmetty, 1995).

The Diagnostic Rhyme Test (DRT)

Based on a set of isolated words to test for consonant intelligibility in initial position. The test consists of 96 word pairs which differ by a single acoustic feature in the initial consonant. The listener hears one word at the time and marks to the answering sheet which one of the two words he thinks is correct. Finally, the results are summarized by averaging the error rates from answer sheets. DRT is widely used and it provides valuable diagnostic information how properly the initial consonant is recognized, variations of DRT for testing middle and final consonants are also in use (Goldstein, 1995).

The Cluster Identification Test (CLID)

A test based on a statistical approach where the test vocabulary is generated for each test sequence separately. The test procedure consists of three main phases: word generator, phoneme-to-grapheme converter and an automatic scoring module. A word generator generates the test material in phonetic representation based on statistical distribution. Certain syllable structures of interest can be defined and tested (Lemmetty, 1995).

The Vowel-Consonant Transition Test (VCTT)

Utilizes nonsense words, mostly transitions between vowels and consonant, and is one of the most commonly used evaluation method for synthetic speech. This method provides high error rates and excellent diagnostic material, especially when an open response set is used. Usually a list of VC, CV, VCV or CVC words is used, but longer words, such as CVVC, VCCV, or CCCVCCC, are sometimes needed. Especially when testing diphone-based systems, longer units must be used to test all diphone combinations (Goldstein, 1995).

Harvard Psychoacoustic Sentence Test

Is based on a closed set of 100 sentences developed to test the word intelligibility in sentence context. The sentences are chosen so that the various segmental phonemes of English are represented in accordance with their frequency of occurrence. Unlike in segmental tests an answer may be correct although some word were missed, especially if meaningful sentences are used. Variations of the test were nonsense sentences are used reduces this effect, but the learning effect is still problematic closed sets are used (Goldstein, 1995).

All of the tests mentioned above are used to assess how a single phoneme or word is recognized, and most intelligibility tests are performed on this level. Intelligibility testing of sentences and paragraphs is also important, since a system with a low comprehension is next to useless. The procedure for intelligibility testing of paragraphs is to let the test subjects listen to a paragraph or more of speech and letting them answer questions concerning what was said. Depending on the application it may also be important to evaluate the pronunciation of proper names since pronunciation of names often does not follow the same rules as pronunciation of other words. Naturalness is much harder to assess than intelligibility. The main problem has to with assessment of pronunciation and prosody. Evaluation of the prosodic features in synthesized speech is probably one of the most challenging tasks in speech synthesis evaluation. Several methods have been developed to evaluate speech quality in general and some of these methods are also suitable to measure the naturalness of synthetic speech, but more pronunciation and prosody oriented evaluation metrics are dearly needed.

Mean Opinion Score (MOS)

This is probably the simplest method of evaluating speech quality in general. It is also suitable for overall evaluation of synthetic speech. MOS is a five level scale from bad (1) to excellent (5), it is also known as ACR (Absolute Category Rating). The listener's task is simply to assess the overall quality of the speech on a discrete scale (Goldstein, 1995).

SAM Prosodic Form Test

Diagnostic evaluation of parameters used to characterise specific intonation contours, regardless of the context in which they appear. Meaningful words of varying segmental complexity in terms of phonological vowel length and syllable structure. Several items are included for each level of complexity, with accurate stress placement. Items are placed in a short, neutral carrier phrase. Each item should be generated with the contours the system can produce in addition to a reference monotone. The assessment of naturalness is done by naive subjects using magnitude estimation (Karlsson, 1998).

SAM Overall Quality Test

Adequacy evaluation of overall quality aspects, particularly acceptability, intelligibility, and naturalness for longer stretches of speech. Eight lists of 20 meaningful sentences of varying syntactic structures and length are used. Each aspect of speech is rated by a different group of subjects. When rating acceptability, it is recommended that application specific speech materials are presented to users. The ratings are based on two sentences each time (Karlsson, 1998).

ITU-TS p.85 Overall Quality Test

Adequacy evaluation of overall quality aspects, Speech samples of between 10 and 30 seconds, adapted to the application are used in the test. It is recommended that a degraded human reference is included. The test subjects are asked to rate the quality on eight categorical estimation scales, they are also asked about the message content (ITU-TS Recommendation P.85, 1993; Karlsson, 1998).

From the overview given here it should be obvious that synthesized speech can be evaluated in many ways, although most are subjective and requires listeners. All methods give some kind of information on the speech quality, but it is easy to see that there is no test to give the one and only correct data (Goldstein, 1995). Perhaps the most suitable way to test a speech synthesizer is to select several methods to assess each feature separately. For example using segmental, sentence level, prosody, and overall tests together provides lots of useful information, but is on the other hand very time-consuming. The test methods must be chosen carefully and it is thus important to consider in advance what kind of data that is needed and why (Lemmetty, 1995).

7. Conclusions

As we have seen there is a large difference between how speech recognition and speech synthesis is evaluated. The reason for this is partially that the tasks that are being evaluated are very different, and partially that the goals with the evaluations have been very different. In one sense, the strategies applied for evaluation of speech synthesis lacks what the strategies applied for evaluation of speech recognition has, and vice versa. For speech recognition there is a need for formative evaluation methodologies that offer more insight into what happens inside the speech recognizer, e.g. more glass box oriented approaches, as well as methodologies that tell more about subjective aspects of the speech recognition. For speech synthesis there is a need for more formative evaluation methodologies that offer more insight into how different speech synthesis methods rate against each other, which implies more black box oriented approaches, as well as methodologies that tell more about the objective aspects of the speech recognition. For both recognition and synthesis there is also a need for more automated evaluation strategies so that formative evaluation can be performed often during development. All in all, there is still a lot more to do in the area of evaluation. The experiences of evaluation so far are however encouraging for further work as they have shown that the efforts put into evaluation often also yields a marked improvement in performance.

References

Baptist, L. and Seneff, S. (2000). Genesis-II: A Versatile System for Language Generation in Conversational System Applications. *Proceedings of ICSLP*, Beijing, China, October 2000, 271 – 274.

Blomberg, M. and Elenius, K. (2000). Automatisk igenkänning av tal. In *Kompendium i talteknologi*. Department of Speech, Music, and Hearing, Royal Institute of Technology (KTH), Stockholm, Sweden.

Carlson, R. and Granström, B. (1997). Speech Synthesis. In *The Handbook of Phonetic Sciences*, W. Hardcastle and J. Laver (Eds.), Blackwell Publishers Ltd, Oxford, 768-788.

Goldstein, M. (1995). Classification of Methods Used for Assessment of Text-to-Speech Systems According to the Demands Placed on the Listener. *Speech Communication*, 16, 225-244

Hirschman, L. and Thompson, S. H. (1997). Overview of Evaluation in Speech and Natural Language Processing. In *Survey of the State of the Art in Human Language Technology*. R Cole et al. (Eds.), Cambridge University Press, Cambridge, 409 – 414.

ITU-TS Recommendation P.85 (1993). A method for subjective performance assessment of the quality of speech voice output devices. International Telecommunications Union. COM 12-R 6-E.

Karlsson, I. (1998). Working Paper on Speech Generation Current Practice. DISC Deliverable D1.3.

Lemmetty, S. (1999). *Review of Speech Synthesis Technology*. Masters Thesis, Department of Electrical and Communications Engineering, Helsinki University of Technology.

McTear, M. F. (2002). Spoken dialogue technology: enabling the conversational interface. *ACM Computing Surveys*, 34(1), 90 – 169.

Sproat, R., Ostendorf, M. and Hunt, A. (1999). The need for increased speech synthesis research. *Report from the 1998 NSF workshop for discussing research priorities and evaluation strategies in speech synthesis*. National Science Foundation.

Zue, V., Cole, R. and Ward, W. (1997). Speech Recognition. In *Survey of the State of the Art in Human Language Technology*. R Cole et al. (Eds.), Cambridge University Press, Cambridge, 3 – 10.

Web Resources

COCOSDA. The International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques. Available at: http://www.slt.atr.co.jp/cocosda/

DARPA. Defense Advanced Research Projects Agency (US). Available at: http://www.darpa.mil/

DISC. The DISC Best Practice Guide to Spoken Dialog Systems (EU). Available at: http://www.disk2.dk/

EAGLES. Expert Advisory Group on Language Engineering Standards (EU). Available at: http://www.ilc.pi.cnr.it/EAGLES96/home.html

ESPRIT. The European information technologies programme (EU). Available at: http://www.cordis.lu/esprit/

LDC. The Linguistic Data Consortium (US). Available at: http://www.ldc.upenn.edu/

NIST. National Institute of Standards and Technology (US). Available at: http://www.nist.gov/