Proceedings of the Institute of Acoustics

EXPECTATIONS FOR ASSESSMENT TECHNIQUES APPLIED TO SPEECH SYNTHESIS

Katherine Morton

University of Essex, Linguistics Department, Colchester

0. INTRODUCTION – THE IMPORTANCE OF ASSESSMENT

Reliable methods of assessing the quality and acceptability of speech output from synthesis systems would be useful to two groups: (a) designers of the hardware and software and (b) users of voice output products. Potential users of synthesizers are currently able to choose among a fairly small range of commercially available products, but this number will grow as the use for voice output systems becomes more obvious. At some point it will be necessary to agree on a method for judging which will be the most suitable device for the task. Researchers designing these new products need a method of assessing which are the best among existing designs, with a view to determining why they are good, and ideally will be able to measure their own designs against a standard.

Although it is not difficult to see the need for assessment, it has not been a simple matter to agree on techniques of evaluation, and a number of questions can be asked: for example, it is not clear whether techniques should be objective or subjective – or a combination of both. Additionally, how much weight should be placed on where the synthesizer will actually be used? For example, the conditions over a telephone line will be different from an airplane cockpit, and yet different in a classroom. Is it essential to measure performance on specific tasks, or is a rating as to general use sufficient? Is it possible to assess potential performance, or should tests be limited to actual performance? Does the cost of the device, or ease of use, i.e. 'value for money' enter into assessment procedures, or are these considerations irrelevant? What methods can be devised to measure quality of output, since quality in general is an elusive concept and difficult to define?

It is clear that assessment is a multi-dimensional problem that will probably not respond to a solution using a single technique. One technique may be ideal for a particular purpose, but another will be more suited to a different purpose. Ideally then we need some clearly stated techniques that can be replicated in the field, and which can be judged by designers and users to be useful and constructive. And it may be the case that ultimately we will only be able to assess by asking the simple question 'which one sounds best to you, under these conditions?' and rely on consensus for labelling the synthesizer suitable or not.

In this paper I intend to look at some type of assessment techniques and suggest possibilities for future work, since there is currently some demand for an evaluation metric for synthesizers.

1. STATING THE PROBLEM

A well-documented experiment was conducted by Holmes in the late 70s (Holmes 1979, 1988). He wished to show that the design of the parallel formant synthesizer, and his implementation of that concept, could more than adequately meet the requirement to produce good intelligible speech output. In this experiment, listeners reported synthetic speech recreated from a careful analysis of natural speech using a parallel formant synthesizer was as good as the original recorded speech in most cases [it should be noted that the material was restricted to the 4kHz bandwidth of the synthesizer]. The reports of this experiment include spectrograms and other displays of both the original natural and the synthetic speech; these demonstrate that the synthesized speech was close to the original.

This particular work is looked at in detail, because it illustrates some of the points I think we need to take into account when designing assessment procedures. I should emphasize this is not a criticism, but is used as a clearly documented starting point for a further consideration of assessment procedures.

A.In this experiment, the purpose was to examine the performance of the synthesizer itself. Holmes did not explicitly have in mind assessment of the analysis procedures which produced the data which drove the synthesizer. The 'hand optimization' of the data was assumed to produce the best performance possible from the synthesizer. Of course it should be remembered that the optimization procedure might have in some unexpected way distorted the analyzed data, and this might have compensated for defects in the synthesizer. However, errors in this part of the procedure were assumed to be small compared with output errors which could arise due to the design of the synthesizer.

B.Holmes was not assessing a synthesis-by-rule system. The purpose was to show that the synthesizer itself could be incorporated within a larger package, knowing it would not detract from the performance of the system as a whole. This illustrates a point to be discussed later; systems can usefully be assessed not only as a whole but also as the separate components which make them up.

C.The tests Holmes conducted were subjective; that is, listeners were asked to decide whether they were listening to synthetic or real speech without having any other information. The test was: could a human being tell the difference? Futher information is available from the spectrograms produced of both types of speech, and a difference between the natural and synthetic speech can be seen. But it is not easy to identify the differences. To enable improvements to Holmes' synthesizer design or the design of other synthesizers it would be necessary to state explicity how these spectrograms differed, and moreover to know objectively what the differences are and identify features in the acoustic output the synthesizer did not replicate.

D. Using the results from Holmes' experiment, is it possible to determine whether another synthesizer can also produce speech output perceived as similar to human speech under certain conditions? The procedure outlined by Holmes can be repeated by supplying the synthesizer with correct acoustic data and optimizing certain features that can be adequately dealt with by the synthesizer hardware. But, given two synthesizers, is it possible to determine which of the two produces speech more like the human original? In this case, the data taken to both must be identical. Since both sets of data were optimized, the test could become a test of how good a human being is at optimizing data as well as a test of the machine, or even a test of how good the synthesizer was at responding to the optimization procedure. And it would not be clear which would be more important in producing the output – the synthesizer hardware, or data optimization; nor would it be clear if the same importance would be similar for both machines. The problem of what is being measured has been compounded.

E.In addition, the two machines might use different techniques for synthesizing the output. For example, nasal formant amplitudes or fricatives might be handled differently. Different machines in these cases might require input data in different forms, or the data itself coded differently. This introduces another variable: the suitability of the method for deriving the original data. Holmes used an analysis technique complementary to his synthesis technique (before the hand optimization); so an analysis procedure complementary to another synthesizer should probably be used for it. But in the total set of procedures of analysis followed by synthesis, how can we determine the divide between analysis and synthesis when comparing two or more different systems? Holmes not only tested the synthesizer's speech output, but also the way the data was derived which drove it.

2. USEFULNESS FOR ASSESSMENT OF THE HOLMES EXPERIMENT

Although it may look as if only the synthesizer itself is being tested, the hardware is part of an overall system. This is important in subjective assessment since a listener is evaluating the goodness of the total system. But it may hold for objective assessment as well. For example, if a synthesizer is required to produce a second formant whose center frequency is 2000Hz, there may be an interaction between parameters within the synthesizer which would not occur in similar combinations with different frequencies on different machines.

Assessment is usually a test of a complete system such as analysis-synthesis systems or text-to-speech systems. The question usually asked is 'Does the speech from A sound better than the speech from B?', and subjects are asked to rank synthesizers within a particular experiment. Although the actual synthesizers may be equally good, the output may sound poor. In these cases, this would probably indicate that the algorithm for converting text to data for driving the synthesizer was creating errors. If the goodness and suitability of the synthesizer had not been established, then it would not be possible to determine the source of the errors unless those errors were known to be typical of sub-components of such systems.

3. SUBJECTIVE AND OBJECTIVE ASSESSMENT

Holmes' tests were essentially subjective: judgements were elicited from listeners and readers were invited to visually compare spectrograms. Although subjective testing is more difficult to control than objective testing, it may be adequate for the purpose. For example, a telephone company may want to determine whether subscribers will react favourably to interacting with a synthesizer. Subscriber reactions are subjective; and a yes/no division might be adequate.

Ideally, we are not concerned only with ranking different systems but also with setting up a standard by which various synthesis systems can be evaluated. The objective of establishing a standard is (a) to remove direct comparison between individual systems (b) to enable quantification of differences between systems in a precise way, and (c) to establish procedures which guarantee replicating results on different occasions (Johnston, 1989). Such standards can be arrived at subjectively or objectively and provide a rating, as distinct from a ranking, of the items being evaluated.

A. Subjective assessment relies on reports of listener response to stimulus items. These opinions vary considerably, but this variation can be evaluated through reliable experimental designs and statistical techniques. What is more difficult is that listeners' judgements depend on perceptual strategies; therefore the design of a useful experiment and interpretation of the results depends on an understanding of that strategy. This has implications for transferring the results of the subjective evaluation to a decision about what is the best synthesizer, or about the best research plan for improvements.

Within psychology, there are competing theories of human speech perception, but most would agree with a model of perception that contained a large top-down element in the process. Looking at synthesis assessment from this point of view, the top-down process plays a major role if listeners are asked to judge or rank the intelligibility of different synthesis systems. For example, in comparing a fairly good system and one which is better, it may be shown that the differences between the systems are statistically insignificant. The top-down model can explain this by suggesting that in listening to the less good system listeners have had to employ more top-down processing than to the better one. And they may be unaware of this process in the short term. However, if the same listeners are asked to compare the two systems using lengthy linguistically complex stimuli they will begin to report that one system induces some fatigue compared with the other. It is well known that fatigue occurs earlier with increased top-down participation in the perceptual process. By measuring fatigue, the diffrence between the two systems may be statistically significant.

It should be possible, in principle, to set up a metric for comparing the intelligibility of synthesis systems based on how much top-down processing had to be brought into the process of listening to them. The more acoustic data, the less top-down processing is necessary. The less acoustic data, the greater the load on the listener's perceptual system. However, obtaining data to support such a perceptual model may not be possible at the moment.

Another aspect of human perception that could be important is categorical perception (Bacri, 1987). Stimuli which vary along a cline are perceived to fall into categories rather than vary continuously. Boundaries can be marked along the stimulus cline; these boundaries are the break-over points between categories. If a parameter output from a synthesis system on such a cline falls just within such a boundary, what will the listener report? Although a listener can judge the correct category, a very slight change in the synthesizer output will cause this boundary to be crossed and the perceived category to shift. A second synthesiser might produce stimuli well within the category boundaries, giving rise to less fatigue; in this case the same slight change will not cause a category shift. Which synthesizer is better? We could say that the second was better, but how would we know that the first was performing near to the category boundaries without devising some kind of test involving progressively changing the stimuli?

The concept of categorical perception is particularly important because the categories themselves and the boundaries on the stimuli clines vary across languages and dialects because the top-down information is different under these different circumstances. For example, if a telephone company wishes to use a particular synthesized dialect or accent in all regions of the country it should choose one which is not operating close to the limits on this parameter. It will therefore need a test for categorical perception accuracy across listeners' different dialects in the evaluation of different synthesizers. The best machine would generate equal fatigue in all listeners whatever their linguistic/phonetic background.

B.Given the difficulties with subjective judgements, it might appear that objective tests would be easier to conduct than subjective assessments. However, testing a synthesis system is not like testing other types of sound systems: for example, tape recorders. Properties such as frequency response, signal-to-noise ratio and distortion can be measured with considerable accuracy. Standards can be set up from these measurements: a studio recorder with a frequency response better than, say, 50Hz–20kHz qualifies for the label 'professional', whereas a domestic machine would meet a lower standard; the rating is derived based on objective measurements. Even in the field of reproduced sound, however, there are those who can hear the difference between tape recorders of identical objective specification and use subjective terms like 'bright' to describe the difference. The objective measurements for tape recorders are based on sinewaves, which occur rarely in the music the machine is required to store and reproduce. Music is created by human beings and contains musical properties in addition to the output of tone generators. This implies that there are features yet to be described and measured for tape recorder evaluation; however, this kind of feature is central to synthesizer evaluation.

The distinguishing features of human speech are extensively described by phonetics. It is well known that no two samples of human speech are the same, even if they are from the same speaker who intends to say the same word in the same way. There is considerable variability not just between utterances but also within a single utterance. This variability centers around the speaker's accuracy at producing speech and it is hypothesized that this occurs because a speaker is speaking to be understood. Speech will improve in direct correlation with the speaker's estimate of the listener's need to draw on top-down information in the perceptual process; thus when a speaker knows that what he has to say may be ambiguous semantically, syntactically or phonologically he will increase the precision of his articulation during a sentence, and sometimes within a word, between phonetic segments. Therefore, the less top-down processing required, the less accurate the speech needs to be. This strategy on the part of the speaker would require a relatively low level use of top-down processing by the listener.

Synthesis systems do not yet incorporate this facility, since this requires knowledge of context which is not yet possible to include in synthesizer design. The result can be a wide swing of top-down processing for the listener over a sentence, and certainly over a period of time. Listeners are not usually aware of their processing, but report that the speech sounds unnatural; this report may contradict the fact that on an objective measurement, each segment and the transitions between them may be perfectly rendered. Thus the objectively good synthesizer, like the objectively good tape recorder, may generate a feeling of unease in the listener.

4. ASSESSMENT IN GENERAL

I should like to separate the general problem of assessment into several smaller problems with associated assessment approaches. (Fourcin *et al.* 1989) The first is to regard synthesizers as complete systems which can be tested on overall performance on the basis of the simple question 'How good does it sound?'. The second approach consists of testing sub-systems for the purpose of making more objective, or combined subject/objective assessment. For example, in Section 2, I suggested separating the synthesizer itself from the algorithm that turns text into signals for driving the synthesizer. But another division which is important with respect to the performance of the driving software lies along linguistically determined lines. For example, we may wish to assess performance of any of the following parameters: (a) segmental rendering, (b) accuracy of of suprasegmentals such as stress, rhythm, intonation, (c) variability of speaking rate, (d) control of intonation, (e) general voice quality, (f) dialectal variation.

If linguistically-based parameters are important, then different types of experiment need to be performed. For example, assessing segmental rendering could be done relatively objectively as could determining the degree of control necessary over fundamental frequency variation, but assessment of general voice quality would probably be subjective.

A. Subjective Procedures
Subjective assessment is based on listeners reports. From the listener's point of view, a test of the synthesis system is based on 'comprehension' and 'goodness', with both being subsumed under 'ease of listening'. Listeners can be asked to make judgements concerning intelligibility of the speech, difficulty in listening or degree of concentration necessary to detect and interpret the signal.

Testing individual speech parameters is possible; for example, intonation, but this assumes that synthesizer parameters can be directly related to perceptual features. In this case the assumption is made that intonation and fundamental frequency correlate. It may be difficult to ensure that all the factors being tested are independent of each other. For example, quality may belong to a different perceptual category than naturalness, although both contribute to judgements concerning overall goodness. Amplitude and duration of segments may contribute toward the perception of stress.

But even with careful experimental design, it is sometimes difficult to decide whether a test is measuring the goodness of the system or whether it is indicating the ability of listeners to decode and report a judgement about a particular parameter. In other words, is it the system's speech output which is being tested or the listener's capabilities. The problem can be increased when the listener is required to make a judgement as to whether the speech is good enough for a particular purpose; the listener must be made aware of the use to which the synthesizer is to be put so that it is clear that the listener's subjective judgements are being made within the right context.

In subjective evaluation experiments it is not clear whether a listener is treating what he hears as degraded natural speech or as some substitute for natural speech (Pisoni and Koehn 1981): it may be that the perceptual system operates in different modes in these two conditions, and this could lead to unclear results. If natural and synthetic speech are processed differently, comparison between various systems will be more difficult to make, since it will not be clear what the listener is actually dealing with. Careful design of a questionnaire for the listener helps to minimize such effects.

The design of the questionnaire is always important in subjective experiments (Cost 209: Sweden). It is essential to ask the right questions of the right subjects and to insure that they understand the questions. Although the experiment will be asking for subjective judgements the terminology used for the experiment should be as precise or as universally understood as possible. For example, a question about whether the speech was distinct, meaning was it slurred, may not be the right question, since not everyone understands generalized terms in the same way.

Many questionnaires ask for listeners' reports by scaling within a range of possible responses (Cost 209: BTRL). The subject is asked to choose the relative strength of a perceived feature within a range predetermined by the experimenter. For example: 'On a scale of one to five, rank the intelligibility of the speech you hear'. Obviously this requires the experimenter to establish properly the most suitable scale for the stimulus items and to be fully aware of the possible expected perceptual categories.

B. Objective Procedures
Objective assessment techniques can take two different routes. They can test whether the synthesizer itself is giving the predicted results according to design specifications: for example, the question can be asked 'is the second formant produced at the required frequency?', or can assess whether the measurable characteristics of natural speech thought to be relevant are accurately replicated. There is an underlying assumption that if the waveform of natural speech is accurately replicated, the speech will sound correct. One test is to look at spectrograms of both natural and synthetic speech. A listener is not involved in such assessment.

On the other hand, if it is not possible to accurately recreate a human speech waveform, what errors actually do matter? For example, is it necessary to accurately render the burst of a plosive when formant bending in the adjacent vowels will provide sufficient information to determine which plosive was intended?

But even in this case, the importance of the accuracy of transitions in general can be questioned. For example, in most text-to-speech systems allophonic units derived from a lookup table are conjoined by rules which interpolate values for each synthesis parameter to ensure a smooth join between segments, thus simulating coarticulation in natural production. The exact method of joining the steady-state segments varies from system to system. A standard way of joining, based on evidence from human speech, would provide a good reference for evaluating the likelihood of a good speech output, but to date it has not been possible to do this within general speech production theory (Tatham, 1985)

Resynthesized speech (Morton 1990) might prove a useful standard for objective assessment. This is a method of coding a natural speech waveform into parameters suitable for driving a synthesizer. Its usefulness lies in the fact that in its coded form this is a representation of natural speech in exactly the same form as the data a synthesis system generates to drive the synthesizer. It is can be compared directly with the results of a text-to-speech algorithm and with the output waveform from human speech.

My own work with resynthesized speech suggestes that transitions can vary considerably before any perceived deterioration in naturalness or quality. Informal experiments, not yet reported, suggest that analyzed human speech can be manipulated to degrade the transitions to be like those produced by a typical synthesis system without loss of perceived accuracy. However in standard synthetic speech systems, poorly rendered transitions are more noticeable. It may well be that in poor segmental synthesis the information provided by transitions is essential, but in resynthesized speech it is assumed that other, more accurate information is made available to the listener from the rest of the waveform and that transitions assume less importance. This is an area we do not yet understand, but it does not detract from the principle that in resynthesized speech we may have some yardstick by which to assess synthesizer performance objectively. (See Section 5).

C. Combined Objective and Subjective Procedures

In general, the combined approach to assessment is based on the correlation between measurable synthesizer parameters and reports from listeners on perceived properties of the test items presented.

Psychoacoustic tests based on known acoustic stimuli are common. One method to check on pronunciation has been tried; the technique involves making reference to a notion of correct pronunciation. This means setting up a standard and mapping deviations from it. It is not a generally successful approach, since what is considered correct varies across listeners. Although the acoustic characteristics of the stimulus items are known, it is not established how representative they are.

At the segmental level, tests such as the diagnostic rhyme test (DRT) are sometimes used (Fourcin *et al.* 1989). These tests rely on consonant confusion arising from contrasting individual segments, with the purpose of testing category perception. But although a listener reports a difference, this difference is not necessarily linguistically relevant. The listener might also report a category shift subjectively, but the objective difference may be so slight that it is difficult to identify within the synthesizer parameter being tested. This problem was referred to above under considerations of perceptual theories.

Because of the degree to which speech perception requires top-down processing, the usefulness of experiments with nonsense words can be questioned. Information from lexical or semantic knowledge of the listener may aid interpreting speech signal in an unknown way. Since synthetic speech is bound to be defective somehow, especially from a text-to-speech system, the listener's tendency to use top-down information may be increased. In this way the listener may push what he hears into an existing category in the case of some nonsense words and perceive them as though they were distorted versions of real words. Other words, more perceptually distant from existing words, may be treated by the listener as genuine nonsense words. It would be difficult to control for this.

It is becoming increasingly important to be able to evaluate suprasegmental phenomena in synthetic speech. In fact, from the point of view of naturalness of the output, prosodic effects can add information which is equal in relevance to segmental rendering. For example, the sentence 'The red light is flashing' can be spoken in many different styles, all of which communicate different types of information.

As yet there are no recognized useful tests in this area (Pols and Boxelaar 1986). We have speech output from synthesis systems generally judged to be monotonous, and machine-like. Again, the question of setting up normative values arises, as does the problem of using nonsense words or sentences in this dimension. It is often not appreciated that the majority of descriptions of intonation, for example, in linguistics are at the phonological level and that their mappping to physical effects at the phonetic lelvel is not completely understood.

5. A LANGUAGE DEPENDENT ASSESSMENT REFERENCE – APIR

Some effort is going into determining a reference standard that is language independent. This is because such a standard automatically allows objective measurements and accurate comparison between systems. But because we are measuring a simulation of part of a language encoding system, we are dealing with assessment of the simulation of a subjective element as well. It is worth looking at the possibility of setting up a reference system dependent on language itself. The proposal is for a measure called the Acoustic Phonetic Information Reference (APIR)

As mentioned above, phonetic and phonological theories claim that speech production and perception are complementary processes. We assume that perception occurs by reference to top-down linguistic knowledge as part of the process of decoding an incoming speech signal. The speech signal triggers the action of this top-down information during the perceptual process. At the phonetic and phonological levels some of this information consists of the listener's knowledge of the phonetic and phonological production processes. In this view, knowledge of speech production is accessed in the decoding process.

In speech production, the speaker makes on-going adjustments to rate of delivery, rhythm, and precision of articulation interacting with the speaker's predictive model of the difficulty of decoding the speech by the listener. The speaker attempts to optimize the output to be complementary with what his/her knowledge of the listener's decoding processes. For example, when speaking to a listener who does not appear to understand the language very well, speakers normally slow down their speech rate, speak louder, and more precisely. It is thought, within this model, that the speaker is trying to minimize the listener's perceptual load. This load is the amount of top-down processing which has to become part of the perceptual processing.

Without the speaker's awareness of the degree to which he appears to be understood, the perceptual load could vary continuously and considerably. Segments and prosodic effects which are ambiguous increase the load; effects which are unique and unambiguous decrease the load.

A listener is aware, often by fatigue or boredom, if the interaction between production and perception process is acceptable. Speech sounds unnatural if it does not contain constant variations just described. No speech synthesis system yet built attempts to simulate this property of human speech production, so no system is designed to collaborate with the listener in optimizing the cognitive load in perception.

As speech synthesis systems beome better at segmental and suprasegmental rendering, and continue to sound unnatural, it becomes increasingly necessary to find a good metric for naturalness. Good segmental and suprasegmental rendering is adequate for short phrases, but the same pattern becomes fatiguing and sounds more unnatural over long periods of speech. This may be because the perceptual load is increasing; natural speech aims to level out wide swings in perceptual loading.

Designers of synthesis systems are now facing the problem of naturalness, because their systems look as if they are good at producing what seems objectively to be a good replication of the human speech waveform. But it is necessary to have a workable definition of what constitutes naturalness and how to measure it. There are other dimensions to naturalness such as adding emotion (Morton 1990) which need descriptions and indices. But for the plain message, I am proposing working out a ratio between information and processing effort as a concept which might be developed into a measurement of naturalness for synthesis systems which can produce a fairly good speech output already.

Since natural speech contains the necessary listener-oriented variability for optimizing the loading, it follows that parametrically analyzed human speech will also contain that variability. It is possible to calculate the difference between parametrically analyzed human speech and a synthesized version of the same utterance based on the same speaker, an index of how defective the synthetic speech is in terms of the variability for that speaker. The index can be called the acoustic phonetic information reference. This difference expresses the proportional relation between the amount of relevant acoustic information provided by the synthetic speech signal and the amount of information it is necessary for the listener to provide. It can show to what extent the synthesis system fails to generate the variability included by a human speaker to optimize perceptual loading on the listener.

The difference would state that the smaller the number, the more accurate the acoustic information; in this case, the listener would need to supply less information for ease of listening. The greater the number, the less accurate the acoustic information; this implies the listener would need to supply more top-down information.

This difference would be useful only when the general quality of synthesis has gone beyond a threshold in terms of accuracy. Two carefully set up modern synthesis systems may sound quite natural for individual words or phrases. But on longer phrases, sentences or paragraphs listeners report unease; a measure could be provided by the difference.

In conclusion, gathering objective measurements and deriving invariant features for setting up standards is extremely difficult, because we are dealing with physical quantitites that are encodings of language. Language itself is the result of cognitive behavior, and as yet we do not have descriptions of this behavior. Linguistic descriptions are about the structure of language, but not about the cognitive processing that has gone on before producing it. For this reason, it is difficult in assessment to be exactly sure of what is being tested. Perhaps the most fruitful area for development is along lines which constitute a combination of both the subjective and objective approaches.

I would like to make the following acknowledgements:

1. British Telecom Research Laboratories (Martlesham Heath) for a research contract into Synthesis Assessment (1990 – on-going), and to R.D. Johnston, B. Surtees, W. Miller and G. Walker for discussions on assessment.

2. G. Garrihy, A. O'Connor, and N. Sifakis at Essex University for research asssistance.

REFERENCES

Bacri, N. (1987) Perceptual spaces and the identification of natural and synthetic sentences, in *Proceedings of International Congress of Phonetics Sciences,* Tallin)

COST209 – Swedish contribution (1987) Subjective quality assessment of synthetic speech, in *Working Party X11/3*

COST209 – BTRL contribution (1988) Methods for subjective determination of transmission quality, in *Study Group X11 Expert's Group on Speech Quality,* Document SQ

Fourcin, A., Harland,G., Barry, W., and Hazan, V. (1989) *Speech Input and Output Assessment.* Chichester: Ellis Horwood Ltd.

Holmes, J.N. (1979) Synthesis of natural-sounding speech using a formant synthesizer, in *Frontiers of Speech Communication Research* (B. Lindblom and S. Ohman, eds). London: Academic Press, pp. 275-285

Holmes, J.N (1988) *Speech Synthesis and Recognition.* Wokingham: VanNostrand Reinhold (UK)

Johnston, R.D. (1989) Speech input/output assessment, in *European Research Project COST209: Supplement* (F.Lundin, ed). Report DUR 12023 EN, pp.228-235

Morton, Katherine (1990) Naturalness in synthetic speech , in *Proceedings of the Institute of Acoustics* Vol 12, Part 10 pp. 125-132

Pisoni,D.B. and Koehn,E (1981) Some comparison of inteligiblity of synthetic and natural speech at different speech-to-noice ratios. In *Research on Speech Perception, Progress Report 7,* Indiana University

Pols, LC.W. and Boxelaar, G.W. (1986) as quoted in Fourcin *op.cit.*

Tatham, M.A.A. (1984) Towards a cognitive phonetics. *Jounmal of Phonetics* 12, pp.37-47