4

The COST 258 Signal Generation Test Array

Gérard Bailly
Institut de la Communication Parlée, UMR-CNRS 5009
INPG and Université Stendhal, 46, avenue Félix Viallet, 38031 Grenoble Cedex 1, France
bailly@icp.inpg.fr

Introduction

Speech synthesis systems aim at computing signals from a symbolic input ranging from a simple raw text to more structured documents, including abstract linguistic or phonological representations such as are available in a concept-to-speech system. Various representations of the desired utterance are built during processing. All these speech synthesis systems, however, use at least a module to convert a phonemic string into an acoustic signal, some characteristics of which have also been computed beforehand. Such characteristics range from nothing – as in hard concatenative synthesis (Black and Taylor, 1994; Campbell, 1997) – to detailed temporal and spectral specifications – as in formant or articulatory synthesis (Local, 1994), but most speech synthesis systems compute at least basic prosodic characteristics, such as the melody and the segmental durations the synthetic output should have.

Analysis-Modification-Synthesis Sytems (AMSS) (see Figure 4.1) produce intermediate representations of signals that include these characteristics. In concatenative synthesis, the analysis phase is often performed off-line and the resulting signal representation is stored for retrieval at synthesis time. In synthesis-by-rule, rules infer regularities from the analysis of large corpora and re-build the signal representation at run-time.

A key problem in speech synthesis is the modification phase, where the original representation of signals is modified in order to take into account the desired prosodic characteristics. These prosodic characteristics should ideally be reflected by covariations between parameters in the entire representation, e.g. variation of the open quotient of the voiced source and of formants according to F_0 and intensity, formant transitions according to duration changes etc. Contrary to synthesis-by-rule systems, where such observed covariations may be described and implemented (Gobl and Chasaide, 1992), the ideal AMSS for concatenative systems

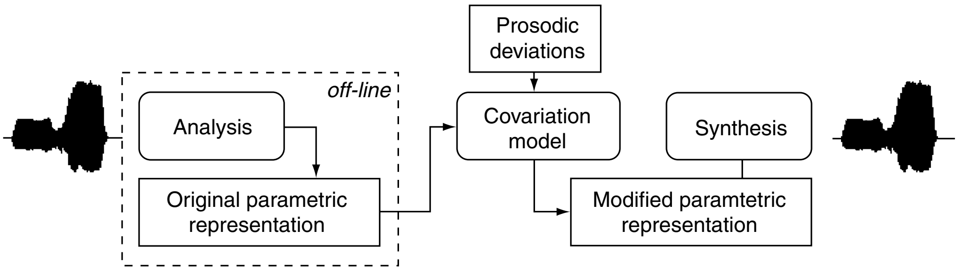

Prosodic deviations
off-line
Analysis
Original parametric representation
Covariation model
Synthesis
Modified paramtetric representation

Figure 4.1 Block diagram of an AMSS: the analysis phase is often performed off-line. The original parametric representations are stored or used to infer rules that will re-build the parametric representation at run-time. Prosodic changes modify the original parametric representation of the speech signal, optimally taking covariation into account

exhibit intrinsic properties – e.g. shape invariance in the time domain (McAulay and Quatieri, 1986; Quatieri and McAulay, 1992) – that guarantee an optimal extrapolation of temporal/spectral behaviour from a reference sample. Systems with a large inventory of speech tokens replace this requirement by careful labelling and a selection algorithm that minimises distortion.

The aim of the COST 258 signal generation test array is to provide benchmarking resources and methodologies for assessing all types of AMSS. The benchmark consists in comparing the performance of AMSS on tasks of increasing difficulty: from the control of a single prosodic parameter of a single sound to the intonation of a whole utterance. The key idea is to provide reference AMSS, including the coder that is assumed to produce the most natural-sounding output: a human being. The desired prosodic characteristics are thus extracted from human utterances and given as prosodic targets to the coder under test. A server has been established to provide reference resources (signals, prosodic description of signals) and systems to (1) speech researchers, for evaluating their work with reference systems; and (2) Text-to-Speech developers, for comparing and assessing competing AMSS. The server may be accessed at the following address: http://www.icp.inpg.fr/cost258/evaluation/server/cost258_coders.

Evaluating AMSS: An Overview

The increasing importance of the evaluation/assessment process in speech synthesis research is evident: the Third International Workshop on Speech Synthesis in Jenolan Caves, Australia, had a special session dedicated to Multi-Lingual Text-to-Speech Synthesis Evaluation, and in the same year there was the First International Conference on Language Resources and Evaluation (LREC) in Grenada, Spain. In June 2000 the second LREC Conference was held in Athens, Greece. In Europe, several large-scale projects have had working groups on speech output evaluation including the EC-Esprit SAM project and the Expert Advisory Group on Language Engineering and Standards (EAGLES). The EAGLES handbook already provides a good overview of existing evaluation tasks and techniques which are described according to a taxonomy of six parameters: subjective vs. objective measurement, judgement vs. functional testing, global vs. analytic assessment, black box vs. glass box approach, laboratory vs. field tests, linguistic vs. acoustic. We will discuss the evaluation of AMSS along some relevant parameters of this taxonomy.

Global vs. Analytic Assessment

The recent literature has been marked by the introduction of important AMSS, such as the emergence of TD-PSOLA (Hamon *et al.*, 1989; Charpentier and Moulines, 1990) and the MBROLA project (Dutoit and Leich, 1993), the sinusoidal model (Almeida and Silva, 1984; McAulay and Quatieri, 1989; Quatieri and McAulay, 1992), and the Harmonic + Noise models (Serra, 1989; Stylianou, 1996; Macon, 1996). The assessment of these AMSS is often done via 'informal' listening tests involving pitch or duration-manipulated signals, comparing the proposed algorithm to a reference in preference tests. These informal experiments are often not reproducible, use *ad hoc* stimuli[1] and compare the proposed AMSS with the authors' own implementation of the reference coder (they often use a system referenced as TDPSOLA, although not implemented by Moulines' team). Furthermore, such a global assessment procedure provides the developer or the reader with poor diagnostic information. In addition, how can we ensure that these time-consuming tests (performed in a given laboratory with a reduced set of items and a given number of AMSS) are incremental, providing end-users with increasingly complete data on a system's performance?

Black Box vs. Glass Box Approach

Many evaluations published to date either involve complete systems (often identified anonymously by the synthesis technique used, as in Sonntag *et al.*, 1999) or compare AMSS within the same speech synthesis system (Stylianou, 1998; Syrdal *et al.*, 1998). Since natural speech – or at least natural prosody – is often not included, the test only determines which AMSS is the most suitable according to the whole text-to-speech process. Moreover, the AMSS under test do not always share the same properties: TD-PSOLA, for example, is very sensitive to phase mismatch across boundaries and cannot smooth spectral discontinuities.

Judgement vs. Functional Testing

Pitch or duration manipulations are usually limited to simple multiplication/division of the speech rate or register, and do not reflect the usual task performed by AMSS of producing synthetic stimuli with natural intonation and rhythm. Manipulating the register and speech rate is quite different from a linear scaling of prosodic parameters. Listeners are thus not presented with plausible stimuli and judgements can be greatly affected by such unrealistic stimuli. The danger is thus

[1] Some authors (see, for example, Veldhuis and Yé, 1996) publishing in *Speech Communication* may nevertheless give access to the stimuli via a very useful server http://www.elsevier.nl:80/inca/publications/store/5/0/5/5/9/7 so that listeners may at least make their own judgement.

to move towards an aesthetic judgement that does not involve any reference to naturalness, i.e. that does not consider the stimuli to have been produced by a biological organism.

Discussion

We think that it would be valuable to construct a check list of formal properties that should be satisfied by any AMSS that claims to manipulate basic prosodic parameters, and extend this list to properties – such as smoothing abilities, generation of vocal fry, etc. – that could be relevant in the end user's choice. Relevant functional tests, judgement tests, objective procedures and resources should be proposed and developed to verify each property.

These tests should concentrate on the evaluation of AMSS independently of the application that would employ selected properties or qualities of a given AMSS: coding and speech synthesis systems using minimal modifications would require transparent analysis-resynthesis of natural samples whereas multi-style rule-based synthesis systems would require highly flexible and intelligible signal representation (Murray *et al.*, 1996). These tests should include a natural reference and compete against it in order to fulfil one of the major goals of speech synthesis, which is the scientific goal of COST 258: improving the naturalness of synthetic speech.

The COST 258 proposal

We here propose to evaluate each AMSS on its performance of an appropriate prosodic *transplantation*, i.e. performing the task of modifying the prosodic characteristics of a source signal in order that the resulting synthetic signal has the same prosodic characteristics as a target signal. We test here not only the ability of AMSS to manipulate prosody but to answer questions such as:

- Does it perform the task in an appropriate way?
- Since manipulating some prosodic parameters such as pitch or duration modifies the timbre of sounds, is the resulting timbre acceptable or more precisely close to the timbre that could have been produced by the reference speaker if faced with the same phonological task?

This suggests that AMSS should be compared against a natural reference, in order to answer the questions above and to determine if the current description of prosodic tasks is sufficient to realise specific mappings and adequately carry the intended linguistic and paralinguistic information.

Description of tasks

The COST 258 server provides both source and target signals organised in various tasks designed to test various abilities of each AMSS. The first version of the server includes four basic tasks:

- *pitch control*: a speaker recorded the ten French vowels at different heights within his normal register.

- *duration control*: most AMSS have difficulty in stretching noise: a speaker recorded short and long versions of the six French fricatives in isolation and with a neutral vocalic substrate.
- *intonation*: AMSS should be able to control melody and segmental durations independently: a speaker recorded six versions of the same sentence with different intonation contours: a flat reference and five different modalities and prosodic attitudes (Morlec *et al.*, forthcoming).
- *emotion*: we extend the previous task to emotional prosody in order to test if prosodic descriptors of the available signals are sufficient to perform the same task for different emotions.

In the near future, a female voice will be added and a task to assess smoothing abilities will be included. AMSS are normally language-independent and can process any speech signal given an adequate prosodic description that could perhaps be enriched to take account of specific time/frequency characteristics of particular sounds (see below). Priority is not therefore given to a multi-lingual extension of the resources.

Physical resources

The server supplies each signal with basic prosodic descriptors (see Figure 4.2). These descriptors are stored as text files:

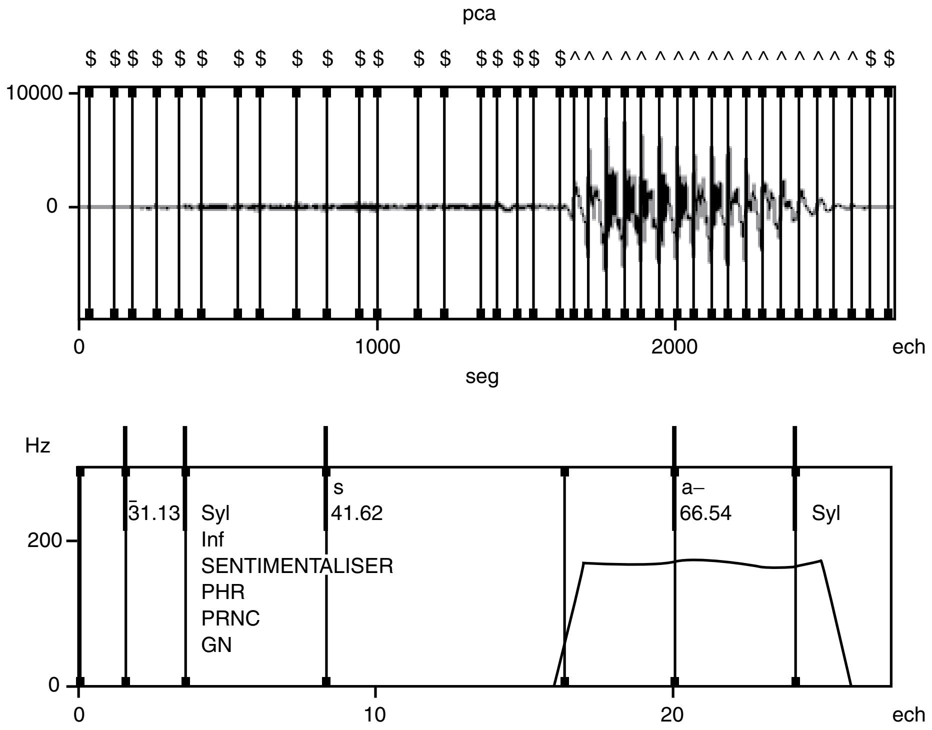


Figure 4.2 Prosodic descriptors of a sample signal. Top: pitch marks; Bottom: segmentation

- Segmentation files (extension .seg) contain the segment boundaries. Short-term energy of the signal (dB) at segment 'centres' is also available.
- Pitch mark files (extension .pca) contain onset landmarks for each period (marked by ^). Melody can thus be easily computed as the inverse of the series of periods. Additional landmarks have been inserted: burst onsets (marked by !) and random landmarks in unvoiced segments or silences (marked by $).

All signals are sampled at 16 kHz and time landmarks are given in number of samples. All time landmarks have been checked by hand.[2]

The Rules: Performing the Tasks

Each AMSS referenced in the server has fulfilled various tasks all consisting in transplanting prosody of various target samples onto a source sample (identified in all tasks with a filename ending with NT). In order to perform these various transplantation tasks, an AMSS can only use the source signal and its prosodic descriptors, the target descriptors.

A discussion list will be launched in order to discuss what additional prosodic descriptors that can be semi-automatically determined should be added to the resources.

Evaluation Procedures

Besides providing reference resources to AMSS developers, the server will also gather and propose basic methodologies to evaluate the performance of each AMSS. In the vast majority of cases, it is difficult or impossible to perform mechanical evaluations of speech synthesis, and humans must be called upon in order to evaluate synthetic speech. There are two main reasons for this: (1) humans are able to produce judgements without any explicit reference and there is little hope of knowing exactly how human listeners process speech stimuli and compare two realisations of the same linguistic message; (2) speech processing is the result of a complex mediation between top-down processes (*a priori* knowledge of the language, the speaker or the speaking device, the situation and conditions of the communication, etc.) and signal-dependent information (speech quality, prosody, etc.). In the case of synthetic speech, the contribution of top-down processes to the overall judgement is expected to be important and no quantitative model can currently take into account this contribution in the psycho-acoustic models of speech perception developed so far.

However, the two objections made above are almost irrelevant for the COST 258 server: all tests are made with an actual reference and all stimuli have to conform to prosodic requirements so that no major qualitative differences are expected to arise.

[2] Please report any mistakes to the author (bailly@icp.inpg.fr).

Objective vs. Subjective Evaluation

Replacing time-consuming experimental work with an objective measurement of an estimated perceptual discrepancy between a signal and a reference thus seems reasonable but should be confirmed by examining the correlation with subjective quality (see, for example, the effort in predicting boundary discontinuities (Klabbers and Veldhuis, 1998).

Currently there is no objective measure which correlates very well with human judgements. One reason for this is that a single frame only makes a small contribution to an objective measure but may contain an error which renders an entire utterance unacceptable or unintelligible for a human listener. The objective evaluation of prosody is particularly problematic, since precision at some points is crucial but at others is unimportant. Furthermore, whereas objective measures deliver time-varying information, human judgements consider the entire stimulus. Although gating experiments or online measures (Hansen and Kollmeier, 1999) may give some time-varying information, no comprehensive model of perceptual integration is available that can directly make the comparison of these time-varying scores possible.

On the other hand, subjective tests use few stimuli – typically a few sentences – and are difficult to replicate. Listeners may be influenced by factors other than signal quality especially when the level of quality is high. They are particularly sensitive to the phonetic structure of the stimuli and may not be able to judge the speech quality for foreign sounds (see, however, the discussion on the extension of the server to other languages on p. 000). Listeners are also unable to process 'speech-like' stimuli.

Distortion Measures

Several distortion measures have been proposed in the literature that are supposed to correlate with speech quality (Quackenbush *et al.*, 1988). Each measure focuses on certain important temporal and spectral aspects of the speech waveform and it is very difficult to choose a measure that perfectly mimics the global judgement of listeners. Some measures take into account the importance of spectral masses and neglect or minimise the importance of distortions occurring in spectral bands with minimal energy (Klatt, 1982). Other measures include a speech production model, such as the stylisation of the spectrum by LPC.

Instead of choosing a single objective measure to evaluate spectral distortion, we chose here to compute several distortion measures and select a compact representation of the results that enhances the differences among the AMSS made available.

Following proposals made by Hansen and Pellom (1998) for evaluating speech enhancement algorithms, we used three measures: the Log-Likelihood ratio measure (LLR), the Log-Area-Ratio measure (LAR), and the weighted spectral slope measure (WSS) (Klatt, 1982). The Itakura-Saito distortion (IS) and the segmental signal-to-noise ratio (SNR) used by Hansen and Pellom were discarded since the temporal organisation of these distortion measures was difficult to interpret.

We will not evaluate the temporal distortion separately since the task already includes timing constraints – which can also be enriched – and temporal distortions will be taken into account in the frame-by-frame comparison process.

Evaluation

As emphasised by Hansen and Pellom (1998), the impact of noise on degraded speech quality is non-uniform. Similarly, an objective speech quality measure computes a level of distortion on a frame-by-frame basis. The effect of modelling noise on the performance of a particular AMSS is thus expected to be time-varying (see Figure 4.3). Although it is desirable to characterise each AMSS by its performance on each individual segment of speech, we performed a first experiment using the average and standard deviation of distortion measures for each task performed by each AMSS and evaluated by the three measures LAR, LLR and WSS, excluding comparison with reference frames with an energy below 30 dB.

Each AMSS is thus characterised by a set of 90 average distortions (3 distortion measures × 15 tasks × 2 characteristics (mean, std)). Different versions of 5 systems (TDPICP, c1, c2, c3, c4) were tested: 4 initial versions (TDPICP0,[3] c1_0, c2_0, c3_0, c4_0) processed the benchmark. The first results were presented at the Cost 258 Budapest meeting in September 1997. After a careful examination of the results, improved versions of three systems (c1_0, c2_0, c4_0) were also tested.

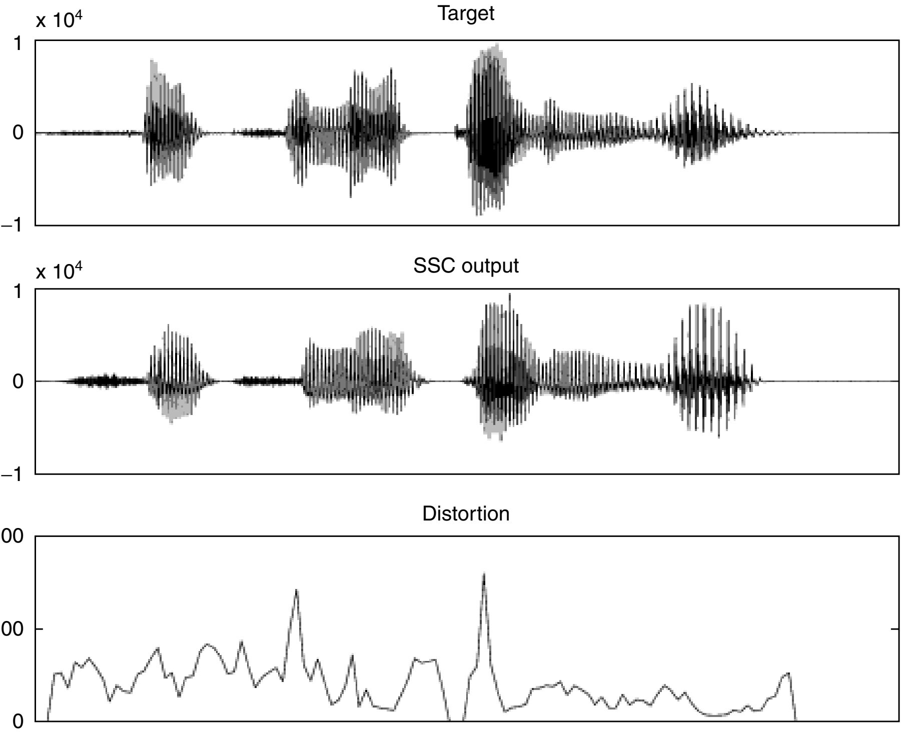


Figure 4.3 Variable impact of modelling error on speech quality. WSS quality measure versus time is shown below the analysed speech signal

[3] This robust implementation of TDPSOLA is described in (Bailly *et al.*, 1992). It mainly differs from Charpentier and Moulines (1990) in its windowing strategy that guarantees a perfect reconstruction in the absence of prosodic modifications.

We added four reference 'systems': the natural target (ORIGIN) and the target degraded by three noise levels (10 dB, 20 dB and 30 dB).

In order to produce a representation that reflects the global distance of each coder from the ORIGIN and maximises the difference among the AMSS, this set of 9 × 90 average distortions was projected onto the first factorial plane (see Figure 4.4) using a normalised principal component analysis procedure. The first, second and third components explain respectively 79.3%, 12.2% and 5.4% of the total variance in Figure 4.4.

Comments

We also projected the mean characteristics obtained by the systems on each of the four tasks (VO, FD, EM, AT) considering the others null. Globally, all AMSS correspond to a SNR of 20 dB. All improved versions resulted in bringing systems closer to the target. This improvement is quite substantial for systems c1 and c2, and demonstrates at least that the server provides the AMSS developers with useful diagnostic tools. Finally, two systems (c1_1, c2_1) seem to outperform the reference TDPSOLA analysis-modification-synthesis system.

The relative placement of the noisy signals (10 dB, 20 dB, 30 dB) and of the tasks (VO, FD, EM, AT) shows that the first principal component (PC) correlates with the SNR whereas the second PC correlates with the ratio between voicing/noise distortion – explained by the fact that FD and VO are placed at the extreme and that a 10 dB SNR has a lower ordinate than the higher SNRs. Distortion measures used here are in fact very sensitive to formant mismatches and when they are drowned in noise, the measures increase very rapidly. We would thus expect that systems c2_0 and c3_0 had an inadequate processing of unvoiced sounds, which is known to be true.

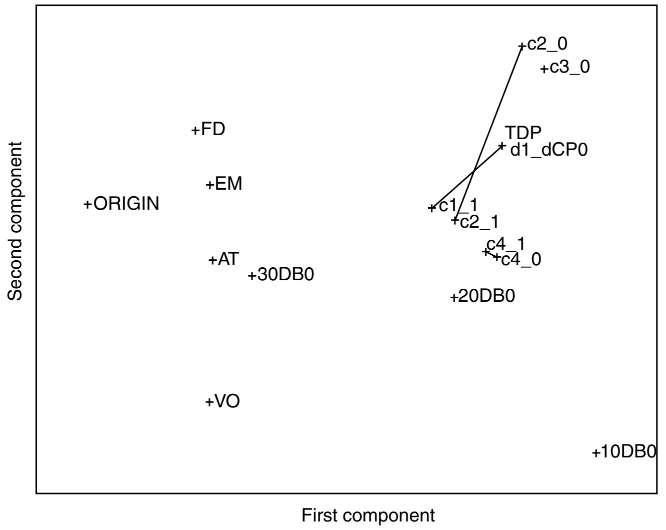
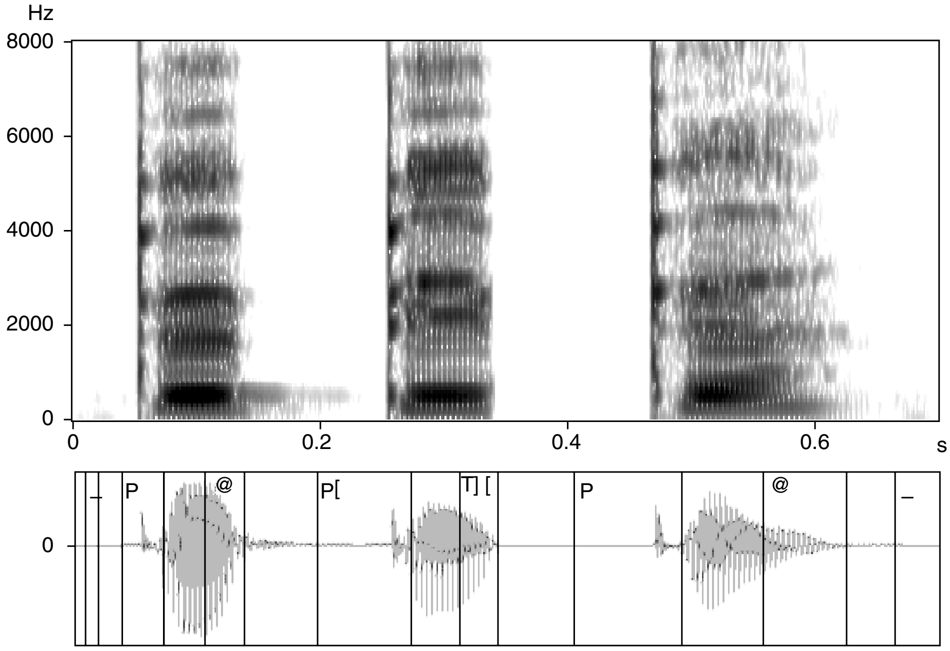
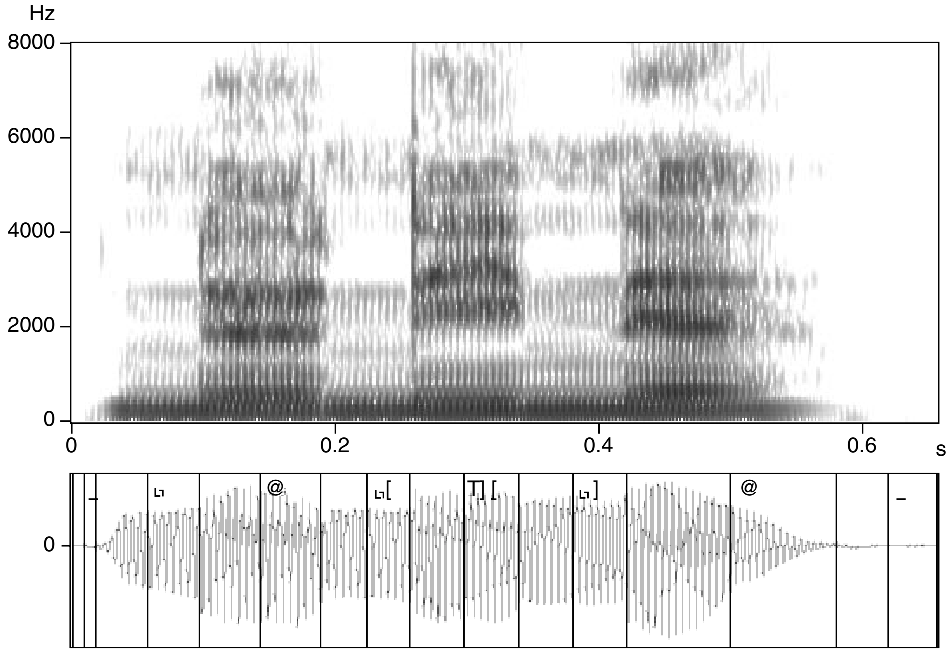


Figure 4.4 Projection of each AMSS on the first factorial plane. Four references have been added: the natural target and the target degraded by 10, 20 and 30 dB noise. c1_1, c2_1, c4_1 are improved version of respectively c1_0, c2_0, c4_0 made after a first objective evaluation


Hz
8000
6000
4000
2000
0
0
0.2
0.4
0.6
s
0
_
P
@
P[
T] [
P
@
_


Hz
8000
6000
4000
2000
0
0
0.2
0.4
0.6
s
0
_
⌞⌝
@
⌞⌝[
T] [
⌞⌝]
@
_

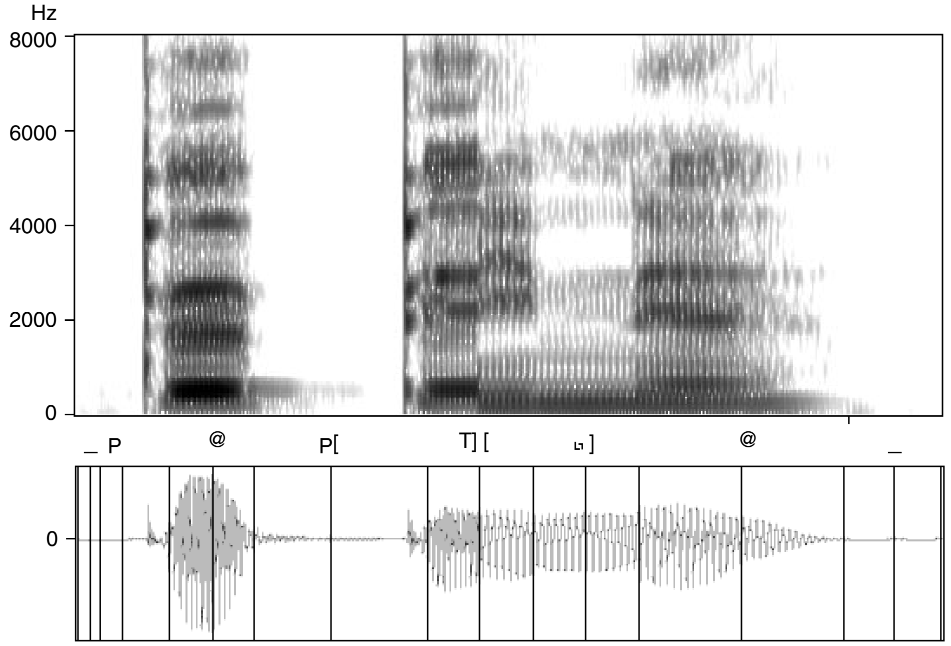


Figure 4.5 Testing the smoothing abilities of AMSS. Left: the two source signals [pɘpipɘ] and [nɘninɘ] Right: the hard concatenation of two signals at the second vocalic nuclei with an important spectral jump due to the nasalised vowel that AMSS will have to smooth

Conclusion

The Cost 258 signal generation test array should become a helpful tool for AMSS developers and TTS designers. It provides AMSS developers with the resources and methodologies needed to evaluate their work against various tasks and results obtained by reference AMSS.[4] It provides TTS designers with a benchmark to characterise and select the AMSS which exhibits the desired properties with the best performance.

The Cost 258 signal generation test array aims to develop a check list of the formal properties that should be satisfied by any AMSS, and extend this list to any parameter that could be relevant in the end user's choice. Relevant functional tests should be proposed and developed to verify each property. The server will grow in the near future in two main directions: we will incorporate new voices for each task – especially female voices – and new tasks. The first new task will be launched to test smoothing abilities, and will consist in comparing a natural utterance with a synthetic replica built from two different source segments instead of one (see Figure 4.5).

[4] We expect to inherit very soon the results obtained by the reference TD-PSOLA implemented by Charpentier and Moulines (1990).

Acknowledgements

This work has been supported by Cost 258 and ARC-B3 initiated by AUPELF-UREF. We thank all researchers who processed the stimuli of the first version of this server, in particular Eduardo Rodriguez Banga, Darragh O'Brien, Alex Monaghan and Miguel Gascuena. A special thanks to Esther Klabbers and Erhard Rank.

References

Almeida, L.B. and Silva, F.M. (1984). Variable-frequency synthesis: An improved harmonic coding scheme. *IEEE International Conference on Acoustics, Speech, and Signal Processing* (pages 27.5.1–4.). San Diego, USA.

Bailly, G., Barbe, T., and Wang, H. (1992). Automatic labelling of large prosodic databases: Tools, methodology and links with a text-to-speech system. In G. Bailly and C. Benoît, (eds) *Talking Machines: Theories, Models and Designs* (pp. 323–333). Elsevier B.V.

Black, A.W. and Taylor, P. (1994). CHATR: A generic speech synthesis system. *COLING-94*, Vol. II, 983–986.

Campbell, W.N. (1997). Synthesizing spontaneous speech. In Y. Sagisaka, N. Campbell, and N. Higuchi (eds), *Computing Prosody: Computational Models for Processing Spontaneous Speech* (pp. 165–186). Springer Verlag.

Charpentier, F. and Moulines, E. (1990). Pitch-synchronous waveform processing techniques for text-to-speech using diphones. *Speech Communication*, *9*, 453– 467.

Dutoit, T. and Leich, H. (1993). MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, *13*, 435– 440.

Gobl, C. and Chasaide. N. (1992). Acoustic characteristics of voice quality. *Speech Communication*, *11*, 481–490.

Hamon, C., Moulines, E., and Charpentier, F. (1989). A diphone synthesis system based on time domain prosodic modification of speech. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, *1*, 238–241.

Hansen, J.H.L. and Pellom, B.L. (1998). An effective quality evaluation protocol for speech enhancement algorithms. *Proceedings of the International Conference on Speech and Language Processing*, *6*, 2819–2822.

Hansen, M. and Kollmeier, B. (1999). Continuous assessment of time-varying speech quality. *Journal of the Acoustical Society of America*, *105*, 2888–2899.

Klabbers, E. and Veldhuis, R. (1998). On the reduction of concatenation artefacts in diphone synthesis. *Proceedings of the International Conference on Speech and Language Processing*, *5*, 1983–1986.

Klatt, D.H. (1982). Prediction of perceived phonetic distance from critical-band spectra: A first step. *IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 1278–1281). Paris, France.

Local, J. (1994). Phonological structure, parametric phonetic interpretation and natural-sounding synthesis. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition* (pp. 253–270). Wiley and Sons.

McAnley, R.J. and Quatieri, T.F. (1986). Speech analysis-synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *ASSP-34(4)*, 744–754.

Macon, M.W. (1996). unpublished PhD thesis, Georgia Institute of Technology.

Morlec, Y., Bailly, G., and Aubergé, V. (forthcoming) Generating prosodic attitudes in French: Data, model and evaluation. *Speech Communication*.

Murray I.R., Arnott J.L., and Rohwer, E.A. (1996). Emotional stress in synthetic speech: Progress and future directions. *Speech Communication*, *20*, 85–91.

Quackenbush, S.R., Barnwell, T.P., and Clements, M.A. (1988). *Objective Measures of Speech Quality*. Prentice-Hall.

Quatieri, T.F. and McAulay, R.J. (1992). Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing*, *40*, 497–510.

Serra X. (1989). *A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic plus Stochastic Decomposition*. PhD thesis, Stanford University, CA.

Sonntag, G.P., Portele, T., Haas, F., and Köhler, J. (1999). Comparative evaluation of six German TTS systems. *Proceedings of the European Conference on Speech Communication and Technology*, *1*, 251–254. Budapest.

Stylianou, Y. (1996). *Harmonic plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis, École Nationale des Télécommunications, Paris.

Stylianou, Y. (1998). Concatenative speech synthesis using a harmonic plus noise model. *ESCA/COCOSDA Workshop on Speech Synthesis* (pp. 261–266). Jenolan Caves, Australia.

Syrdal, A.K, Möhler, G., Dusterhoff, K., Conkie, A., and Black, A.W. (1998). Three methods of intonation modeling. *ESCA/COCOSDA Workshop on Speech Synthesis* (pp. 305–310). Jenolan Caves, Australia.

Veldhuis, R. and Yé, H. (1996). Time-scale and pitch modifications of speech signals and resynthesis from the discrete short-time Fourier transform. *Speech Communication*, *18*, 257–279.