The Cost258 Signal Generation Test Array

Gérard Bailly[1], Eduardo R. Banga[2], Alex Monaghan[3], Erhard Rank[4]

[1] ICP - UMR CNRS n°5009, INPG & U3, 46 av. Félix Viallet, 38031 Grenoble CEDEX, France
[2] ETSI Telecomunicacion, Campus Universitario, Universidad de Vigo, 36200 Vigo, Spain
[3] National Centre for Language Technology, Dublin City University, Dublin 9, Ireland
[4] INTHF, Vienna University of Technology, Gusshausstrasse 25/E389, A-1040 Vienna, Austria
Cost258 evaluation server : `www.icp.inpg.fr/cost258/evaluation/server/cost258_coders.html`

Abstract

This paper describes a benchmark for Analysis-Modification-Synthesis Systems (AMSS) that are back-ends of all concatenative speech synthesis systems. After introducing the motivations and principles underlying this initiative, we present here a first anonymous objective evaluation comparing the performance of 5 such AMSS.

1. Introduction

Most Text-to-Speech (TTS) systems generate speech signals by concatenating natural speech segments of which some characteristics are smoothed at concatenation points and further modified according to prosodic instructions (see fig. 1) computed by the TTS system in order to encode appropriate linguistic/paralinguistic information. Systems using a large database of segments can minimise or eventually suppress this modification step by an adequate selection of segments (Campbell, 1997; Klabbers and Veldhuis, 1998). But even in this case, it seems however difficult to ensure an homogeneous speech quality over the whole database and to avoid using an AMSS, at least for scaling or smoothing the intermediate parametric representation of the speech signals.

For speech synthesis, this modification step is crucial. Whereas speech coding just consists of quantisation, more drastic processing is required when modifying the entire temporal/spectral parametric structure. The modification of a parameter such as the fundamental frequency (F0) of the speech signal correlates with covariations in the entire parametric representation, e.g., with characteristics of the glottal waveform and formants, a phenomenon widely studied (Gobl and Chasaide, 1992) and used in synthesis-by-rule systems. When AMSS do not exploit such intelligible parameters, these covariations should be implicit, i.e., elicited by more global properties of the AMSS. The most common property is shape invariance: maintaining the global shape of the waveform in the vicinity of excitation instants. Shape invariance is common to most AMSS such as TD-PSOLA (Charpentier and Moulines, 1990) or sinusoidal synthesisers (Quatieri and McAulay, 1989) ... This preservation of signal properties is however not sufficient to guaranty the coherence of the synthetic temporal/spectral structure and it seems necessary for now to have a comparative evaluation of AMSS that checks if they are able to reproduce the modification observed in natural speech.

This paper describes a benchmark developed within the Cost258 action that provides resources, reference AMSS and objective methods for evaluating new AMSS.

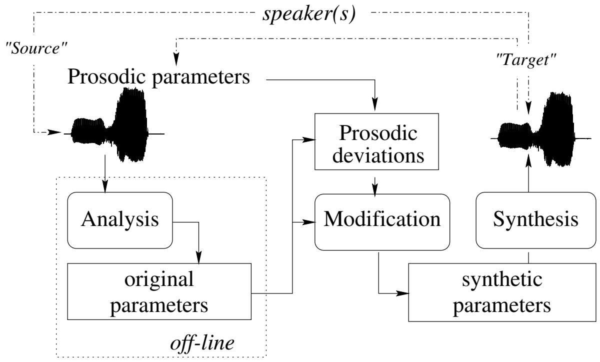

Figure 1: Block diagram of a AMSS: a parametric representation of natural signals is modified according to prosodic instructions computed by the TTS. The evaluation process consists of prosodic transplantation tasks where source and target signals are uttered by the same speaker.

2. Evaluation of AMSS: an ill-posed problem

Thanks to the emergence of statistical methods in speech synthesis, the freely available generic tools (Dutoit et al., 1996; Taylor et al., 1998) and multilingual resources, building a TTS is no longer necessarily the fruit of the strenuous work of speech scientists. It is thus very important to provide means to evaluate the intrinsic properties of each module of this giant generic "lego". Few "glass box" evaluation procedures have been proposed and tested: most evaluations challenge complete TTS systems (van Santen et al., 1998) - often identified anonymously by the AMSS they use (see for example (Sonntag et al., 1999)) - or compare AMSS within the same TTS system architecture (Dutoit, 1994; Stylianou, 1998; Syrdal et al., 1998). Apart from the fact that most evaluation procedures do not include a reference scale such as natural reference signals or at least natural prosody, it is impossible to establish an evaluation grid because the properties of AMSS regarding the tasks they are confronted with are so different. For example, TD-PSOLA is very sensitive to concatenation problems, due to the difficulty of smoothing both phase and amplitude spectra, but can preserve signals (Böeffard and Violaro, 1994) that are difficult to analyse with other parametric AMSS with a great precision. On the other hand, new proposals of AMSS are often accompanied with informal evaluation tests using unrealistic tasks (such as constant tempo or F0 manipulations) on ad hoc stimuli [1]. And hence the listeners are biased towards an aesthetic judgement without any reference to the expected properties and performance of the AMSS. And these properties are diverse: a TTS system that makes use of a large sound database in order to avoid large prosodic modifications will require a transparent analysis-synthesis process with a high compression whereas multi-style synthesis systems making largely use of various models of contextual speech variability will require a versatile AMSS enabling severe modifications of the parametric representation of signals.

2.1. Identifying tasks

The evaluation grid we propose relies on the identification of elementary tasks coders are confronted with and from which noticeable differences among AMSS behaviours are expected. Such tasks should include the manipulation of elementary prosodic parameters such as melody, tempo and intensity, but also the modification of spectral quality for diverse aims such as spectral smoothing, multi-style synthesis, or voice transformation.

The AMSS should also be confronted with different sound classes: manipulating the F0 of voiced fricatives poses the problem of the synchronisation of friction noise and glottal cycle (Hermes, 1991); lengthening unvoiced sounds points out the problem of the definition and processing of speech frames in most AMSS.

2.2. From elementary to complex tasks

As mentioned in the introduction, these manipulations of the parametric representation are not independent and should be made synergetically to structure the discourse. The evaluation grid should thus cover not only the AMSS but also the module that computes the diverse modifications to be performed by the AMSS.

2.3. Reference coder(s)

Recent AMSS are almost always compared to TD-PSOLA. Besides the fact that currently no reference implementation of this popular AMSS is available and that everybody refers to a customised implementation, we should not forget that we have another reference system at hand: the human speaker; which could be asked to perform these various tasks. Although some authors claim that AMSS could beat human performance for very special tasks, we are still far from this objective for most tasks.

The main originality of the test array is thus to consider human performance as THE ultimate target and confront AMSS with various prosodic transplantation tasks involving couples of natural SOURCE and TARGET stimuli (see fig. 1) pronounced by the same speaker for most tasks or by different speakers for the future speaker transformation tasks.

[1]Note the Speech Communication initiative that enables the authors (see for example (Veldhuis and Hé, 1996)) to give a free access to their stimuli via a web server: `www.elsevier.nl:80/cas/tree/store/specom/free`.

3. The test array

The Cost258 Signal Generation Test Array provides resources, reference AMSS and objective methods for evaluating AMSS.

3.1. Resources

3.1.1. Sounds

The source and target signals are obtained by instructing the speaker either explicitly - using for example a textual or verbal description of a situation that would elicit a particular intonation - or implicitly - using natural/synthetic stimuli suggesting directly (by reiteration) or indirectly (see for example how Barbosa (Barbosa, 1994) obtained statements at five distinct speech rates using synthetic questions) the task to be performed.

3.1.2. Descriptors

The server provides reference prosodic descriptors for all signals in order to avoid gross analysis errors and to hide target signals. These descriptors include at least pitch-marks and a phonemic segmentation. They can be enriched by other "intelligible" descriptors such as spectral slope, estimation of parameters of the voice source…

4. The server in use

4.1. Present resources

All tasks consist of transplanting a neutral (monotonous) version of a sound, a word or a sentence (indicated in the server by NT appended to the filename) towards various versions of the some content. The NT source utterances approximate an ideal concatenation system which will have solved all coarticulatory problems before prosodic manipulation (see for example the preprocessing of MBROLA (Dutoit, 1994)).

All signals have been sampled at 16kHz, segmented and pitch-marked semi-automatically. In addition, the centres of realizations of each phoneme have been marked and their short-term energy added to the set of descriptors.

The speakers have fulfilled four types of tasks :

V0 (F0 control) speakers recorded the ten French vowels at different F0 apart from their normal register;

FD (duration control) speakers recorded short and long versions of the six French fricatives in isolation and with a neutral vocalic substrate;

AT (intonation control) speakers recorded 6 versions of the same sentences with different intonation contours: a flat reference and five different modalities and prosodic attitudes;

EM (prosody control in emotional speech) extension of the AT corpus to emotional prosody.

4.2. AMSS studied

Several AMSS performed the various tasks: a well tried version of TD-PSOLA implemented at ICP (Bailly et al., 1992), four initial AMSS (c1_0,c2_0,c3_0,c4_0) and an improved version of three of them (c1_1,c2_1,c4_1) resulting from an initial evaluation presented in a Cost258 meeting in Budapest. Three of these systems use the Harmonic+Noise Model (HNM) (Bailly, 1999; O'Brien and Monaghan, 1999; Banga et al., 1997) and one use Residual-Excited Linear Prediction (RELP) (Rank and Pirker, 1998). Three additional references were generated by adding white noise with different signal-to-noise ratios (SNR) to the source stimuli: 30dB, 20dB and 10dB. The server gives an interactive access to all source and target stimuli and to synthetic signals.

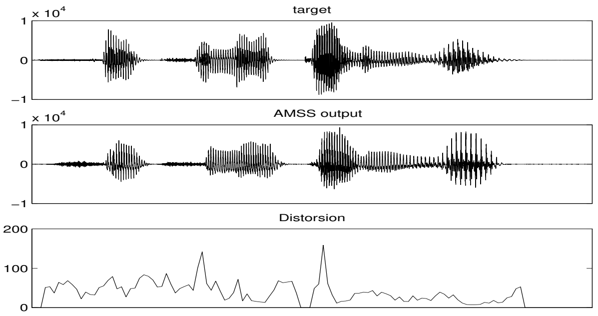

Figure 2: Time-varying distortion computed by the WSS measure.

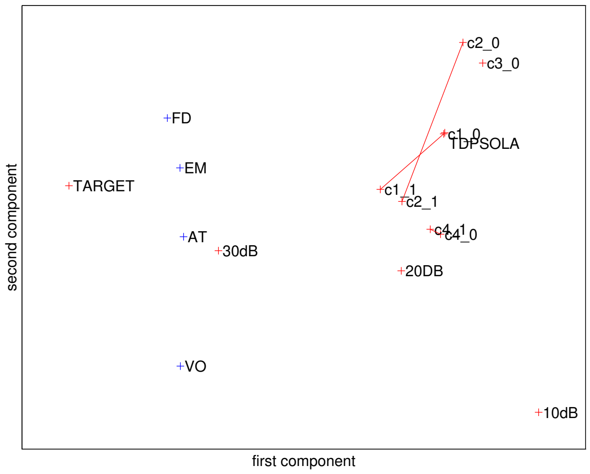

Figure 3: Projection of 11 signal sets (5 AMSS + 3 new versions + 3 SNR) on the first factorial plane obtained by applying a PCA of the set of the average distortions they obtain on the transplantation tasks. Reference systems (TARGET,corpora) have been projected afterwards.

5. Objective Evaluation

The project also aims at providing basic methodologies for the objective evaluation of AMSS. Numerous works try to correlate Mean Opinion Scores (MOS) with objective measurements of estimated distortions of signals, e.g., for predicting the perceptual discomfort elicited by listening to spectral discontinuities (Ding et al., 1998; Klabbers and Veldhuis, 1998) or assessing the performance of speech enhancement algorithms (Hansen and Pellom, 1998).

5.1. Distortion measures

Several measures have been proposed in the literature that are supposed to correlate with speech quality (Quackenbush et al., 1988). Each measure focusses on certain important temporal and spectral aspects of the speech waveform and it seems very difficult to choose a measure that mimics perfectly the global judgement of listeners. Moreover these measures deliver time-varying information (see fig. 2) that is difficult to correlate with a global judgement [2]: we will thus only consider the global behaviours of distortion measures, i.e., the mean and standard deviation across each utterance.

Instead of choosing a single objective measure to evaluate spectral distortion we choose here to compute several distortion measures and leave the selection of the best combination of results to a Principal Component Analysis (PCA) (see below).

Following proposals made by Hansen and Pellom (Hansen and Pellom, 1998) for evaluating speech enhancement algorithms, we use 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 SNR ratio used by Hansen and Pellom were discarded since the temporal organisation of these distortion measures was difficult to interpret.

5.2. Displaying and interpreting results

Each of the 11 sets of signals (5 AMSS + 3 new versions + 3 SNR) is thus characterised by a set of 90 statistical outcomes (3 distortion measures x 15 tasks (5 pitch scales + 1 duration lengthening + 5 attitudes + 4 emotions) x 2 characteristics (mean, std)). We performed a PCA on this matrix of 11 observations characterised by 90 parameters. The three first principal components explain 78.2%, 11.4%, and 5.6% of the total variance. We projected each AMSS onto the first factorial plane (see fig. 3). We projected also the ideal system with no distortions (TARGET) and 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 20dB. 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. The relative placement of the noisy signals (10dB, 20dB, 30dB) and of the tasks (VO, FD, EM, AT) evidences 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 10dB 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 thus expect that systems c2_0 and c3_0 have an inadequate processing of unvoiced sounds, that is undoubtfully true.

Conclusions and outlook

We invite AMSS developers to submit their systems to the Cost258 server. All resources and technical details can

[2] Note however the nice experiment performed by Hansen and Kollmeier (Hansen and Kollmeier, 1999).

be obtained from the web site. We hope to develop the site towards three directions: (1) enrich the set of prosodic descriptors of the reference signals (2) valorize the different properties of AMSS by the introduction of new tasks such as spectral smoothing or speech quality manipulation, (3) offer new methodologies for objective evaluation and cumulate results of the evaluation experiments performed on subsets of these stimuli (O'Brien and Monaghan, 1999).

Acknowledgements

This work has been supported by Cost 258 and ARC-B3 initiated by AUPELF-UREF.

6. References

Bailly, G., 1999. Accurate estimation of sinusoidal parameters in an harmonic+noise model for speech synthesis. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 3.

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 Gérard Bailly and Christian Benoît (eds.), *Talking Machines: Theories, Models and Designs*. Elsevier B.V., pages 323–333.

Banga, E.R., Mateo, C.G., and Fernández-Salgado, X., 1997. Shape-invariant prosodic modification algorithm for concatenative text-to-speech synthesis. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 1. Rhodes - Greece.

Barbosa, P., 1994. *Caractérisation et génération automatique de la structuration rythmique du français*. Thèse de troisième cycle, Institut National Polytechnique de Grenoble, Grenoble, France.

Böeffard, O. and Violaro, F., 1994. Improving the robustness of text-to-speech synthesizers for large prosodic variations. In *ETRW on Speech Synthesis*. New Paltz - New York.

Campbell, W.N., 1997. Synthesizing spontaneous speech. In Yoshinori Sagisaka, Nick Campbell, and Norio Higuchi (eds.), *Computing prosody: Computational models for processing spontaneous speech*. Springer Verlag, pages 165–186.

Charpentier, F. and Moulines, E., 1990. Pitch-synchronous waveform processing techniques for text-to-speech using diphones. *Speech Communication*, 9(5-6):453–467.

Ding, W., Fujisawa, K., and Campbell, N., 1998. Improving speech synthesis of CHATR using a perceptual discontinuity function and constraints of prosodic modification. In *ETRW Workshop on Speech Synthesis*. Jenolan Caves - Australia.

Dutoit, T., 1994. High quality text-to-speech synthesis : A comparison of four candidate algorithms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1. Adelaide - Australia.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and van der Vrecken, O., 1996. The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proceedings of the International Conference on Speech and Language Processing*, volume 3. Philadelphia - USA.

Gobl, C. and Chasaide, N., 1992. Acoustic characteristics of voice quality. *Speech Communication*, 11(4-5):481–490.

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

Hansen, M. and Kollmeier, B., 1999. Continuous assessment of time-varying speech quality. *Journal of the Acoustical Society of America*, 105(5):2888–2899.

Hermes, D.J., 1991. Synthesis of breathy vowels : Some research methods. *Speech Communication*, 10:497–502.

Klabbers, E. and Veldhuis, R., 1998. On the reduction of concatenation artefacts in diphone synthesis. In *Proceedings of the International Conference on Speech and Language Processing*, volume 5.

Klatt, D.H., 1982. Prediction of perceived phonetic distance from critical-band spectra : a first step. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

O'Brien, D. and Monaghan, A., 1999. Shape invariant pitch modification of speech using a harmonic model. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 3. Budapest - Hungary.

Quackenbush, S., Barnwell, T., and Clements, M., 1988. *Objective measures of speech quality*. Englewood Cliffs, NJ: Prentice-Hall Inc.

Quatieri, T.F. and McAulay, R.J., 1989. Phase coherence in speech reconstruction for enhancement and coding applications. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1.

Rank, E. and Pirker, H., 1998. Realization of prosody in a speech synthesizer for German. In *Proceedings of Konvens'98*. Bonn.

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

Stylianou, Y., 1998. Concatenative speech synthesis using a Harmonic plus Noise Model. In *ESCA/COCOSDA Workshop on Speech Synthesis*. Jenolan Caves, Australia.

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

Taylor, P., Black, A.W., and Caley, R., 1998. The architecture of the FESTIVAL speech synthesis system. In *ESCA/COCOSDA Workshop on Speech Synthesis*. Jenolan Caves, Australia.

van Santen, J.P., Pols, L.C., Abe, M., Kahn, D., Keller, E., and Vonwiller, J., 1998. Report on the third ESCA Workshop evaluation procedure. In *ESCA/COCOSDA Workshop on Speech Synthesis*. Jenolan Caves, Australia.

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