FUTURE WORK IN COST ACTION 277 ON NONLINEAR SPEECH PROCESSING[1]

Gérard Chollet, Marcos Faúndez-Zanuy, Bastiaan Kleijn, Gernot Kubin, Steve McLaughlin, and Bojan Petek

e-mail: chollet@tsi.enst.fr, faundez@eupmt.es, bastiaan@speech.kth.se, g.kubin@ieee.org, sml@ee.ed.ac.uk, bojan.petek@uni-lj.si

COST277 website: http://www.ee.ed.ac.uk/~cost277

ABSTRACT

In this paper we present the progress of the COST action 277 on nonlinear speech processing. The action is focused on four key themes: speech synthesis, speech coding, speech recognition and speech analysis. The purpose of this paper is twofold, to state clearly and publicly the aims of the action for the next three years, and also to attract the interest of researchers and laboratories around the world thus potentially increasing collaborations between them.

1. INTRODUCTION

COST is an intergovernmental framework for European co-operation in the field of scientific and technical research. Since its foundation in 1971, COST has provided a valuable mechanism for coordinating national research activities in Europe. The primary objective of the action is to improve the quality and capabilities of the voice services for telecommunication systems through the development of new nonlinear speech processing techniques. The following countries are currently contributing to COST-277: Austria, Belgium, France, Germany, Ireland, Italy, Portugal, Slovenia, Spain, Sweden, Switzerland and United Kingdom, with the likelihood that Slovakia and Greece will join in the near future.

The action is broken down into 4 working groups each with a focus on one of the key areas of the action; synthesis, recognition, coding and analysis. It should be emphasised that the separation of the working groups is for management convenience and members have interests in several areas and cross-fertilisation between areas is actively encouraged. This paper summarizes the technical working plan for the next three years. The paper is structured as follows, initially we present plans for each of the four working groups in individual sections followed by a short summary and conclusions.

2. SPEECH ANALYSIS

Speech analysis is of central importance to all research areas of speech processing, including speech synthesis, speech and speaker recognition and speech coding. Speech analysis techniques that exhibit noise robustness are of vital importance for successful deployment of modern information and communication technologies. Speech analysis research, however, is highly interdisciplinary, integrating various areas such as, eg, signal processing, acoustics, phonetics, phonology, psychoacoustics and cognition.

The COST 277 speech analysis workgroup aims to address and contribute to the new theoretical and empirical advances in the field, including: the voice source analysis, non-linear enhancement of dysphonic voices, speech enhancement, signal generation techniques for speech synthesis, creation of new tools for speech analysis and recognition, as well as to the multilinguality and portability issues in speech processing. In the following, some of the research avenues mentioned above are elaborated more in detail.

Novel theoretical contributions are expected in the area of voice source analysis that consider analysis and synthesis of phonatory excitation signal by a polynomial waveshaper model [18]. This model, for example, can prove efficient in the synthesis of emotional speech since it enables precise control of the shape of the excitation signal. Experimental work in this area will involve a study of modulation frequency and modulation levels in the context of vocal microtremor analysis. Vocal cycle length perturbation decomposition will be investigated and will include an analysis of jitter and vocal microtremor for normophonic speakers. Furthermore, analysis of the acoustic primitives of phonatory patterns will be performed by a multivariate analysis of the flat vowel spectra for dysphonic voices. Future work is going to include analysis of connected speech instead of sustained vowels, analysis of the pathological instead of the normal tremor, and synthesis of the disordered voices and other timbres.

[1] This work has been supported by the European Commission under the COST Action 277: http://www.ee.ed.ac.uk/~cost277

Advances are expected in non-linear enhancement of dysphonic voices [5]. The plan is to acquire a database of dysphonic speech, perform additional standard voice assessment on about 100 informants using 20 min of speech per speaker. Some recordings will be made in various surroundings and situations (providing a database of 10 speakers with 2 hrs/speaker). This research will also address and develop novel algorithms that aim to improve dysphonic speech by phase-space domain methods. Implementations will involve nonlinear noise reduction techniques. A study of applicability of the nonlinear oscillator model to enhance dysphonic speech will be performed. Additional research avenues in speech enhancement are detailed in [4].

Research relevant to the COST 277 synthesis working group will concentrate on the signal generation techniques for speech synthesis [6]. This research is going to involve an acquisition of a speech database, automatic and semi-automatic determination of contextual indicators and acoustic measures such as the position of a sound within a word or a syllable, pitch determination (F0), F0 cycle variation, amplitude, and segment duration. ANN-based analysis of interdependencies and predictive structures is going to be developed. Identification of acoustic parameters of speech, such as style of speech, emotive state, gender, and speaker individuality will be incorporated into a harmonics-and-noise based signal generation algorithm. Specifically, this research aims to construct an integrated, (semi)-automatic segmentation and analysis system that combines the speech synthesis systems for French and German with the IKP aligner. Parameters will be stored in an SQL database that will provide a basis for extensive explorations of the relationships between the linguistically relevant symbolic and acoustic parameters.

Another line of research aims to address a cross-section of the multilinguality and portability issues in speech processing with the development of human language technologies [15]. Specifically, signal analysis and processing techniques that could potentially help to ameliorate the serious problem of the lack of resources for non-prevalent languages could also be investigated.

3. SPEECH SYNTHESIS

In the speech synthesis workgroup the focus is on developing nonlinear techniques for improved speech synthesis. The primary motivation for looking at nonlinear approaches is that despite the long-standing research showing that "comprehensible speech synthesis is good enough for most folks", really, truly high-quality speech synthesis

(a) has a tremendous impact;

(b) will eventually supplant low-quality speech synthesis, no matter what;

(c) will thus inevitably dominate the scientific and market scene of the future.

For example, concatenative TD speech synthesis methods can be made to sound reasonably good, (and form the basis of most commercial TTS systems), but they are too inflexible, since they require a new data base for each new voice and voice style. There are literally millions of voices and voice styles we would like to model potentially. In addition traditional generation algorithms for synthesis (e.g. formant synthesis) do poorly, because of inadequate modelling of high-frequency material.

However, current non-linear approaches still need work [9,11] and it is in this area that the work in the COST action will concentrate.

Specific work is underway on a variety of approaches;

(a) hidden node neural-network models (HNM);
(b) Non-linear oscillator approaches;
(c) Hybrid solutions;
(d) Local linear embedding techniques.

All of these techniques are focussing on voiced speech to offer more natural speech. More detail on some of the techniques is available at this conference in [11].

4. SPEECH CODING

In speech coding the objective is to convert the speech signal into a bit stream that is devoid of irrelevant and redundant information and to reconstruct a speech signal of the required quality from this bit stream. The efficiency of this coding process will benefit strongly from nonlinear processing techniques.

The removal and reconstruction of irrelevant signal content is determined by the processing performed by the human auditory system and, furthermore, by processing within the brain. This processing of acoustic signals is well-known to be highly nonlinear (e.g., [2,13]). We will explore two main avenues for exploiting current knowledge of the processing of the human auditory system for the coding of the speech signal. First, we intend to create improved perceptual error criteria. Second, we intend to create an invertible model of the human auditory system, thus enabling efficient coding in a perceptual domain. This approach will extend the work of [8].

The removal of redundancy from the signal also benefits strongly from nonlinear processing. In coding linear operators are the cause of artificial sounding speech at low rates, and, conversely, of the high rates required for natural-sounding speech. We will reduce these problems by introducing generic nonlinear techniques such as

thresholding in combination with adaptive frame expansions, extending [7]. More importantly, by creating a structure containing nonlinear operators that allows a more accurate (where accuracy is measured by perception) model of the speech signal, we expect to obtain more efficient speech coding methods. This work will build on [10].

Very low bit rate speech coding (below 400 bps) can be achieved by indexing a memory of segmental units which is shared by the coder and the decoder [1]. This approach makes use of recognition and synthesis techniques which are a common interest for all Working Groups of the COST-277 action.

5. SPEECH & SPEAKER RECOGNITION

Recognition is by essence a nonlinear process. The input speech signal is continuous; the output is symbolic. Depending on the task, the output could be an orthographic text corresponding to what was said, an action to accomplish in response to a request, the detection of key words or other acoustic events, the detection of a change of speakers, the identity of the speaker or the verification of a claimed identity, the detection of a speech pathology, an objective measure of articulation, of intelligibility, or of quality. A recognizer needs some reference data which could either be compared directly with the speech input, or serve as training samples to construct models of the units to recognize.

The time waveform should be analysed to extract parameters which are relevant for comparison. Speech production is a dynamic process. Most of the analysis tools currently being used are based on a short-time stationary assumption which is an oversimplification. More effort should focus on parametric and non-parametic time-frequency distributions [3, 16, 17]. Non-linear predictors may be able to follow the time-dependent trajectories in the feature space and achieve syllable-size segmentation and labeling.

Hidden Markov Models (HMM) have demonstrated their efficiency in capturing some of the variability of speech signals. They have limitations which are studied within the more general framework of Bayesian Networks and graphical models. Dynamic Bayesian networks can model an arbitrary set of variables as they evolve over time. This allows the joint distribution to be represented in a highly factored way [19]. Research must focus on inference techniques of the structure of the network as well as estimation of parameters.

In many cases, there is a mismatch between training and testing conditions: the speaker was not available during training, the environmental or transmission conditions differ. Ideally, acoustic features should be insensitive to this mismatch. High Order Statistics could be exploited to increase robustness [12]. Speech is being transmitted over packet-switched networks (eg: Internet). Some packets may be lost and techniques to maintain high recognition rates need to be studied [14].

Speaker Independent Speech Recognition should be insensitive to inter-speaker variability. Consequently, speech and speaker recognition should use different feature sets. In current practice, Automatic Speaker Recognition systems usually make use of the same features as for Speech Recognition. Further investigations on time-frequency representations and comparative evaluations for both speech and speaker recognition are needed.

6. CONCLUSIONS

It is clear that in the signal processing community nonlinear approaches offer potential significant benefits for many problems. Speech is no different and the COST 277 action is actively exploring a wide variety of techniques for the four key application areas of speech processing. COST is an ideal way of ensuring collaborative European research and anyone who is interested is asked to contact any of the authors for further information or to look at the web site http://www.ee.ed.ac.uk/~cost277.

7. REFERENCES

[1]J. Cernocky, G. Baudoin and G. Chollet. Segmental vocoder - going beyond the phonetic approach. *Proc. ICASSP'98*, pp. 605-608, Seattle, 1998.

[2]T. Dau, D. Püschel and A. Kohlrausch. A quantitative model of the "effective" signal processing in the auditory system. I. Model structure, *J. Acoust. Soc. Am.*, 99(6): 3615 - 3622, 1996.

[3]Y. Grenier. Parametric time-frequency representations. *In: Signal Processing*, J-L. Lacoume, T.S. Durrani and R. Stora (eds), pp. 339-397, North-Holland Physics Publishing, 1985.

[4]M. Faúndez-Zanuy, G. Kubin, W. B. Kleijn, P. Maragos, S. McLaughlin, A. Esposito, A. Hussain, J. Schoentgen. Nonlinear Speech Processing: Overview and Applications. ***International Journal of Control and Intelligent Systems*, Special Issue on Nonlinear Speech Processing Techniques and Applications (in press),** paper available at http://www.ee.ed.ac.uk/~cost277/actapress_final.pdf

[5]M. Hagmüller and G. Kubin. Nonlinear enhancement of dysphonic voices. *Presentation at the COST 277 MCM in Graz*,http://www.ee.ed.ac.uk/~cost277/program_graz_2.pdf April 2002.

[6]E. Keller. E-mail correspondence, April 2002.

[7]W. B. Kleijn. A Frame Interpretation of Sinusoidal Coding and Waveform Interpolation, *Proc. ICASSP'00*, Vol 3:1475-1478, Istanbul, 2000.

[8]G. Kubin and W. B. Kleijn. On Speech Coding in a Perceptual Domain. *Proc. ICASSP'99*, Vol I:205-208, Phoenix, 1999.

[9]I. N. Mann, S. McLaughlin. Synthesising natural-sounding vowels using a nonlinear dynamical model, *Signal processing*, Vol 81, pp 1743-1756, 2001.

[10]P. Maragos, J. Kaiser, and T. F. Quatieri. Energy Separation in Signal Modulations with Application to Speech Analysis, *IEEE Trans. Signal Processing*, 41(10), pp. 3024-3051, October 1993.

[11]S. McLaughlin. Nonlinear Speech Synthesis, *Proceedings of EUSIPCO 2002*, Toulouse.

[12]A. Moreno, S. Tortola, J. Vidal and J.A.R. Fonollosa. New HOS-based parameter estimation methods for speech recognition in noisy environments. *Proc. ICASSP'95*, pp. 429-432, Detroit, 1995.

[13]R. D. Patterson, M. H. Allerhand and C. Giguère. Time-domain modeling of periphal auditory processing: A modular architecture software platform, *J. Acoust. Soc. Am.*, 98(4): 1890 - 1894, 1995.

[14]C. Pelaez-Moreno, E. Parrado-Hernandez, A. Gallardo-Antolin, A. Zambrano-Miranda and F. Diaz-de-Maria. Nonlinear methods for packet loss reconstruction in ASR applications. *Proc. EUSIPCO*, Toulouse, 2002.

*[15]*B. Petek (ed.). Portability Issues in Human Language Technologies, *Proc. LREC 2002 workshop, June 2002. Workshop information available at URL* http://www.lrec-onf.org/lrec2002/lrec/wksh/Portability.html

[16]J. W. Pitton, L.E. Atlas and P.J. Loughlin. Applications of positive time-frequency distributions to speech processing. *In IEEE Trans. on SAP*, Vol. 2, No 4, pp. 554-566, 1994.

[17]J. Rouat, S. Lemieux and A. Migneault. A spectro-temporal analysis of speech based on nonlinear operators. *Proc. ICSLP'92*, Vol. 2:1629-1632, Banff, 1992.

[18]J. Schoentgen. Analysis and synthesis of the phonatory excitation signal by means of a polynomial waveshaper. *Presentation at the COST 277 MCM in Graz*, April 2002. http://www.ee.ed.ac.uk/~cost277/program_graz_2.pdf

[19]G. G. Zweig. Speech Recognition with Dynamic Bayesian Networks. *PhD Thesis*, University of California, Berkeley, 1998.