The COST-277 European Action: An overview

Marcos Faundez-Zanuy (*), Unto Laine, Gernot Kubin, Stephen McLaughlin, Bastiaan Kleijn, Gerard Chollet, Bojan Petek, Amir Hussain.

(*) Escola Universitaria Politècnica de Mataró (Spain)
faundez@eupmt.es

Abstract. This paper summarizes the rationale for proposing the COST-277 "nonlinear speech processing" action, and the work done during these last four years. In addition, future perspectives are described.

1 Introduction

COST-277 is an innovative approach: so far, cost actions where focused on a single application field: Speaker Recognition in Telephony (COST-250), Naturalness of synthetic speech (COST-258) [1], Spoken Language interaction in telecommunication (COST-278), etc. However, there are strong arguments for a global approach, which considers speech processing from a general point of view, rather than focussing on a single topic. Section 2 summarizes the rationale for this general approach and the goals of COST-277. Section 3 summarizes the work done inside the framework of COST-277 and section 4 is devoted to results and future lines.

2 Rationale for a speech processing COST action

The four classical areas of speech processing:

1. Speech Recognition (Speech-to-Text, StT)
2. Speech Synthesis (Text-to-Speech, TtS and Code-to-Speech, CtS)
3. Speech Coding (Speech-to-Code, StC with CtS) and
4. Speaker Identification &Verification (SV)

have all developed their own methodology almost independently from the neighboring areas. (See the white arrows in the Figure 1.)

This has led to a plurality of tools and methods that are hard to integrate. Some of the ideas of COST action were to study the following fundamental questions:

- Are there any parametric, discrete models or representations for speech useful for most or even all of the tasks mentioned?
- What properties should these representations have?
- How can the parameters of these models be estimated automatically from continuous speech?

In Fig. 1 natural, human speech is on the left and its synthetic counterpart on the right. Two main methods to compress the speech information are depicted in the mid-

dle. The “written speech” refers to standard orthography of the language or to phonetic writing. The “coded speech” refers to a plurality of different coding methods and parametric representations.

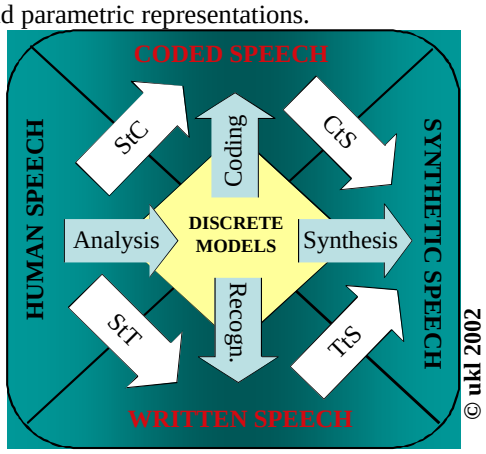

Fig. 1. Classical, separate speech processing areas (white arrows) and an advanced, multifunctional, integrated platform utilizing common discrete models and representations for speech signals.

The coded speech is less compressed and may have high quality whereas the written speech is strongly compressed and without any side information it has lost, e.g., the identity and the emotional state of the speaker.

The simplest codes, like PCM, can be called *one-quality-layer* codes or representations (see Fig. 2). The code is directly related to one and only one quality, attribute or dimensionality, e.g., signal amplitude or sound pressure. These simplest coding methods do not apply models. Model free methods lead to universal coding where the waveform may represent any type of time varying signal: temperature, music, speech etc.

Two-quality-layer codes and representations make the primary separation between source (excitation) and filter (vocal tract and lip radiation). They apply source-filter theory and related models. The filter is typically assumed to be all-pole. All of the possible zeroes of the signal together with the temporal fine structures are modeled by the source (e.g., CELP). These methods may take into consideration the non-uniform frequency resolution of the human auditory system by applying auditory frequency scales (PLP, WLP).

The modeling can be developed further, e.g., by including aspects of articulation and/or related spectral dynamics. These codes can be called *three-quality-layer* codes. The corresponding methods and models are here called “discrete models”. Further, when more complicated structures are found and coded we approach phonetic qualities, descriptions, and codes (IPA). Finally, linguistic qualities and structures are modeled and coded (speech understanding).

The “discrete models” area in the middle of Fig. 1 denotes methods that are scalable and able to produce variable quality, depending on the purpose and capacity of the target system.

Advanced models and methods may be linear, non-linear or combinations of both.

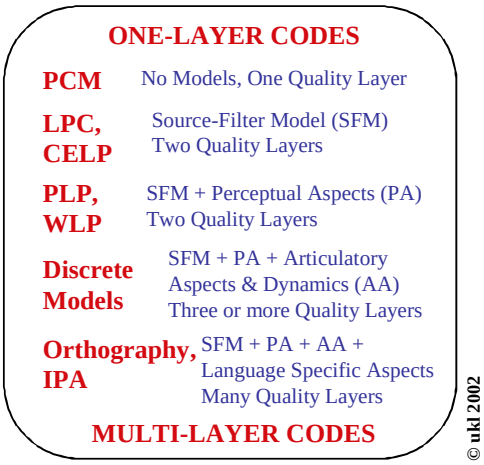

Fig. 2. Different levels of coding and representations of speech signals.

The models and methods should not only help to integrate different speech processing areas, but in addition they should -if possible- posses the following features and reflect:

- Properties of human perception (auditory aspect)
- Properties related to articulatory movements (motoric aspect)
- Inherent features of phonemes or subsegmentals
- Allow mappings between speakers (normalization)
- Robustness: Insensitivity to ambient noise and channel (wireless and packet-based ones) distortions
- Underlying dynamics of the speech signal.

The parametric models capable of reflecting aspects of speech production could help to understand the ”hidden structures” of the speech signal. They could provide tools for more detailed analysis and study on the (acoustical) coding principles of phones or diphones in different contexts of continuous speech.

All this should be of help in understanding the mechanisms of the coarticulation, too. This problem and weakness in current speech recognition schemes should be transformed into power and strength useful in other speech processing areas as well. Phonetic research -especially related to articulatory phonetics and subsegmentals- could benefit of these new tools and methods.

Deeper understanding and efficient modeling the reflections of the speech production mechanism in continuous speech signal and in its phones are in the focus of COST-277.

Source-filter models are almost always part of speech processing applications such as speech coding, synthesis, speech recognition, and speaker recognition technology. Usually, the filter is linear and based on linear prediction; the excitation for the linear filter is either left undefined, modeled as noise, described by a simple pulse train, or described by an entry from a large codebook. While this approach has led to great advances in the last 30 years, it neglects structure known to be present in the speech signal. In practical applications, this neglect manifests itself as an increase in bit rate, a

less natural speech synthesis, and an inferior discriminating ability in speech sounds. The replacement of the linear filter (or parts thereof) with non-linear operators (models) should enable us to obtain an accurate description of the speech signal with a lower number of parameters. This in turn should lead to better performance of practical speech processing applications.

For the reasons mentioned above, there has been a growing interest in the usage of non-linear models in speech processing. Several studies have been published that clearly show that the potential for performance improvement through the usage of non-linear techniques is large.

Motivated by the high potential benefits of this technology, US researchers at well-known universities and industrial laboratories are very active in this field. In Europe, the field has also attracted a significant number of researchers. The COST-277 (Non-linear speech processing) project is the first step towards the creation of a scientific community and the possibility of European collaborative efforts. The initial COST-277 working plan is published in [2].

3 Overview of COST-277 research activities

COST-277 started on June 2001 and will officially finish in June 2005. Section 4.2 contains a list of participating countries. During these last four years, several meetings, workshops and training schools have been organized and articulated the research activities. Four main working groups have been established and worked close together:

1. WG1: Speech Coding.
2. WG2: Speech Synthesis.
3. WG3: Speech and speaker recognition.
4. WG4: Voice Analysis and enhancement.

The main scientific events inside COST-277 action are described in the next subsections.

3.1 Management Committee meetings

The administrative and scientific matters have been discussed in several meetings, whose minutes can be found on the website. The three initial MCM (0, 1, 2) have been organized for the start up of the action, and the remaining ones have included other activities summarized in the next sections.

MCM-0 (pre-inaugural meeting): September, 1999, Budapest (Hungary)
MCM-1 (Inaugural meeting): June, 2001, Brussels (Belgium)
MCM-2 (Unofficial EUROSPEECH meeting): September, 2001, Aalborg (DK)
MCM-3 (Vietri Sul Mare meeting): 6th/7th December, 2001, Vietri Sul Mare (Italy)
MCM-4 (Graz meeting): 11th/12th April, 2002, Graz (Austria)
MCM-5 (Unofficial EUSIPCO'02 meeting): September, 2002, Toulouse (France)
MCM-6 (Edinburgh meeting): 2nd/3rd December 2002, Edinburgh (UK)
MCM-7 (Le Croisic meeting): 20th to 23th May, 2003, Le Croisic (France)
MCM-8 (Laussane meeting): 5th/6th September, 2003, Laussane (Switzerland)

MCM-9 (Limerick meeting) : 15th/16th April, 2004, Limerick (Ireland)
MCM-10 (2nd Vietri sul Mare meeting): 13th-17th September'04
MCM-11 (Barcelona meeting): 19th-22th April'05
MCM-12 (Creete meeting): To be announced.

3.2 Special Sessions organized in signal processing conferences

Five special session have been organized in well-established conferences:

1. International Workshop on Artificial Neural Networks (IWANN'01): Held in June 2001 in Granada (Spain). 3 technical presentations.
2. European Speech Processing Conference: Held in September 2002 in Toulouse (France). 5 technical presentations.
3. International Workshop on Artificial Neural Networks (IWANN'03): Held in June 2003 in Menorca Island (Spain). 4 technical presentations.
4. European Speech Processing Conference: Held in September 2004 in Vienna (Austria). 5 technical presentations.
5. International Conference on Artificial Neural Networks (ICANN'05): To be held in Poland in 2005.

3.3 Workshops

Two main workshops, named NOLISP, have been organized:

1. NOLISP'2003: Held in May 2003 in Le Croisic (France). 32 technical presentations.
2. NOLISP'2005: Held in April 2005 in Barcelona (Spain). 42 technical presentations.

3.4 Training schools

Two training schools have been organized:

1. Vietri Sul Mare (Italy): Held in September 2004. 36 technical presentations [4].
2. Crete (Greece): To be announced

3.5 Short term scientific missions

Two short term missions have been organized:

1. STM-1: Held during 19 June 2003 to 21 June 2003 from Belgium (Brussels) to Laussane (Switzerland): Synthesis of disordered speech and the insertion of voice quality cues into text-to-speech systems.
2. STM-2: from Graz (Austria) to Canada: research on auditory modeling (summer '04)

3.6 EU Framework Programm 6 (FP6) Initiatives

Two expressions of interest were submitted to the European Commission within the FP6 programme:

1. Advanced Methods for Speech Processing (AMSP): http://eoi.cordis.lu/dsp_details.cfm?ID=38399
2. Human Language Technology Portability (HLTport): http://eoi.cordis.lu/dsp_details.cfm?ID=32189

4 Results and future lines

One of the great successes of COST-277 has been the increase of contributions between different countries and other COST actions. This has let to deal and study new research topics, summarized in section 4.3.

4.1 Collaboration with other COST actions

COST-219ter: Accessibility for all to services and terminals for next generation networks

COST-219ter has showed a strong interest on the work "Enhancement of Disordered Male Voices" done by a COST-277 team. Possible future interactions between both COST actions are being studied.

COST-275: Biometric-based recognition of people over the Internet.

Several COST-277 members have attended regular workshops of COST-275 and presented results related to Speaker recognition. In addition, COST-277 has produced the COST-277 database for speaker recognition, which is suitable for the study of new techniques such as:

- Speech signal watermarking for improving the security on remote biometric applications.
- Speech signal bandwidth extension for improving the naturalness of encoded speech.

This database will be available in 2006 [3]. In addition, a joint brochure was disseminated at major conferences, at the beginning of the action.

COST-276: Information and knowledge management for integrated media communication systems

Contacts have been established with COST-276 WG-4 due to the interest on Speech watermarking. However, although COST-276 has interest on speech watermarking, they are more focus on audio (music) watermarking. Thus, COST-277 has a more mature technology for speech, which will be transferred to COST-276.

COST-278: Spoken language interaction in telecommunication

COST-278 members will attend the NOLISP'05 workshop in Barcelona and will present some topics and problems that could be addressed with NL speech processing.

On the other hand, NL speech feature extraction for speech recognition will be presented to COST-278. In addition, a joint brochure was disseminated at major conferences, at the beginning of the action.

4.2 Collaboration between different countries

In order to summarize the different collaborations between institutions, we have just represented in a matrix the collaborations between different countries, made possible thanks to COST-277 action. Next diagram summarizes the inter-country collaborations.

	A	B	CAN	CH	CZ	D	E	F	UK	GR	I	IRL	LT	P	S	SI	SK
A	■	■	■				■	■	■	■	■				■	■	
B	■	■		■			■		■	■	■				■		
CAN	■		■				■	■									
CH		■		■				■									
CZ					■												
D						■											
E	■	■					■	■	■	■	■				■	■	
F	■		■	■			■	■	■		■	■			■	■	
UK	■	■					■	■	■	■	■				■	■	
GR	■	■					■		■	■	■				■		
I	■	■					■	■	■	■	■				■		
IRL							■					■					
LT													■				
P														■			
S	■	■					■	■	■	■	■				■	■	
SI	■	■					■	■	■						■	■	
SK																	■

A shadowed cell means contribution between respective file and row countries (joint publications and/or Short Term Missions).

Country codes

A	Austria
B	Belgium
CAN	Canada
CH	Switzerland
CZ	Czech Republic
D	Germany
E	Spain
F	France
UK	United Kingdom
GR	Greece
I	Italy
IRL	Ireland
LT	Lithuania
P	Portugal

S	Sweden
SI	Slovenia
SK	Slovakia

4.3 Scientific results

A detailed explanation of scientific achievements is beyond the goal of this paper, and can be found in our website www.nolisp2005.org thus, we restrict this section to an enumeration of research activities:

- Analysis and synthesis of the phonatory excitation signal by means of a polynomial waveshaper.
- Modulation frequency and modulation level owing to vocal micro-tremor
- Decomposition of the vocal cycle length perturbations into vocal jitter and vocal microtremor and comparison of their size in normophonic speakers.
- Acoustic primitives of phonatory patterns
- Multivariate Statistical Analysis of Flat Vowel Spectra with a View to Characterizing Disordered Voices
- Relevance of bandwidth extension for speaker identification and verification
- Waveform speech coding using non-linear vectorial predictors.
- SVM-Based Lost Packets Concealment for ASR Applications Over IP
- Space–time representation
- Nonlinear masking of a time–space representation of speech
- Adaptive nonlinear filtering and recognition with neurons
- Nonlinear masking and networks of oscillatory neurons
- Speech structure and masking
- Speech analysis
- What can predictive speech coders learn from speaker recognizers?
- Nonlinear features for speaker recognition
- Speaker recognition improvement using blind inversion of distortions
- Isolating vocal noise in running speech via bi-directional double linear prediction analysis
- On the bandwidth of a shaping function model of the phonatory excitation signal
- Speech signal watermarking: a way to improve the vulnerability of biometric systems

4.4 Future lines

COST-277 will officially finish in June 2005. However, a final event will be held in the last semester of 2005 in the form of a training school. Afterwards, the Nonlinear speech processing community should survive without the European Science Foundation founding. In NOLISP'2003 it was stated the interest for keep on working on this topics, and to stay close to the speech processing community, rather than nonlinear processing groups (image, communications, etc.). Probably, a good idea

would be the establishment of an ISCA Special Interest Group (SIG) on Nonlinear speech processing.

You can keep informed by looking at the website!.

References

1. Keller, E. et al (Eds.) Improvements in Speech Synthesis. John Wiley 2002.
2. Chollet G., Faundez-Zanuy M., Kleijn B., Kubin G., McLaughlin S., and Petek B. “A description of the cost-277 nonlinear speech processing action working-plan”. Vol. III pp.525-528 EUSIPCO’2002, Toulouse.
3. Faundez-Zanuy M., Hagmüller M., Kubin G., Nilsson M., Kleijn W. B., “The COST-277 speech database”. ISCA Workshop NOLISP’2005 Barcelona.
4. G. Chollet, A. Esposito, M. Faundez-Zanuy, M. Marinaro Eds., “Advances in nonlinear speech processing and applications”. LNCS 3435, Springer Verlag 2005.

Acknowledgement

This work has been supported by FEDER and the Spanish grant MCYT TIC2003-08382-C05-02, and European COST action 277 “Non-linear speech processing”. B. Petek acknowledges grant 3311-04-837052 from Ministry of Education, Science and Sport of Republic of Slovenia.