CONTROL OF VOICE QUALITY FOR EMOTIONAL SPEECH SYNTHESIS

Carlo Drioli*, Fabio Tesser^, Graziano Tisato*, Piero Cosi*, Enrico Marchetto°

* Istituto di Scienze e Tecnologie della Cognizione, Sezione di Fonetica e Dialettologia, ISTC-CNR, Padova.
^Centro per la Ricerca Scientifica e Tecnologica, ITC-IRST, Trento.
°Università di Padova, Dipartimento di Ingegneria dell'Informazione, Padova.

[drioli,tisato,cosi]@pd.istc.cnr.it
tesser@irst.itc.it
enrico.m82@libero.it

SUMMARY

Speech production in general, and emotional speech in particular, is characterized by a wide variety of phonation modalities. Voice quality, which is the term commonly used in the field, has an important role in the communication of emotions through speech, and nonmodal phonation modalities (soft, breathy, whispery, creaky, for example) are commonly found in emotional speech corpora.

In this paper, we describe a voice synthesis framework that allows to control a set of acoustic parameters which are relevant for the simulation of nonmodal voice qualities. The set of controls of the synthesizer includes standard controls for duration and pitch of the phonemes, and additional controls for intensity, spectral emphasis, fast and slow variations of the duration and amplitude of the waveform periods (for voiced frames), frequency axis warping for changing the formant position, and aspiration noise level.
Some guidelines are given to combine these signal transformations in the aim of reproducing some nonmodal voice qualities, including soft, loud, breathy, whispery, hoarse, and tremulous voice. It is also discussed how these voice qualities characterize the emotional speech .
The system described here is based on the FESTIVAL speech synthesis framework and on the MBROLA diphone concatenation acoustic back-end. We also address the possibility of including affective tags in the input text to be converted. To this aim, FESTIVAL was provided with the support for the use of affective tags through ad-hoc mark-up languages (APML/VSML), and for driving the extended MBROLA synthesis engine through the generation of voice quality controls. The control of the acoustic characteristics of the voice signal is based on signal processing routines applied to the diphones before the concatenation step. Time-domain algorithms are used for the cues related to pitch control, whereas frequency-domain algorithms, based on FFT and inverse-FFT, are used for the cues related to the short-term spectral envelope of the signal.

1. INTRODUCTION

The transmission of emotions in speech communication is a topic that has recently received considerable attention. Automatic speech recognition (ASR) and text-to-speech (TTS) synthesis are examples of popular fields in which the processing of emotions can have a substantial impact and can improve the effectiveness and naturalness of the man-machine interaction. Many of the researches in the field have emphasized the importance of prosodic features (e.g., speech rate, F0 and intensity contours, F0 range) and the importance of the voice quality in the rendering of different emotions in verbal communication(Gobl & Chasaide, 2003; Johnstone &. Scherer, 1999; Ladd *et alii*, 1985). In TTS technologies, voice processing algorithms for emotional speech synthesis have been mainly focusing on the control of phoneme duration and pitch, which are the principal parameters conveying the prosodic information. On the side of voice quality transformations for speech synthesis, some recent studies have addressed the exploitation of source models within the framework of articulatory synthesis to control the characteristics of voice phonation (d'Alessandro & Doval, 1998; Gobl & Chasaide, 2003). When using the phoneme concatenation approach, one possible solution is to record a different database of phonemes for each emotion (Schröder & Grice, 2003).
The aim of this paper is to describe our solutions adopted for the control of voice quality within a TTS framework. The paper is organized as follows. In Section 2 the voice material is introduced and the principal acoustic cues considered are described. In Section 3 we describe the signal processing approach for phoneme processing, for transforming the voice quality characteristics of speech. In Section 4 a mark-up language for emotional TTS synthesis is described, and in Section 5 the audiovisual integration for facial animation is illustrated.

2. EMOTIONAL SPEECH CORPORA AND ACOUSTIC ANALYSIS

A male University student pronounced two phonological structures 'VCV, corresponding to two feminine proper names: "Aba" /'aba/ and "Ava" /'ava/, simulating, on the basis of appropriate scenarios, six emotional states: anger (A), joy (J), fear (F), sadness (SA), disgust (D) and surprise (SU), apart from the neutral one (N), corresponding to a declarative sentence. This 14 words set was repeated many times in random order. A description of the voice material in terms of acoustic cues commonly related to emotions can be found in (Drioli *et alii*., 2003).

3. SPEECH SYNTHESIS

The most conceptually simple way of generating speech is to concatenate short segments of pre-recorded speech, selected from a database of units (usually words, syllables, demisyllables, phonemes, diphones, and triphones). The speech units are encoded and stored in a database. Previous attempts to model voice quality in such a framework have been based on recording separate diphone databases for different levels of vocal efforts or voice qualities (Schröder, Grice, 2003). However, this approach presents serious drawbacks: excess of memory occupation, complexity of the voice design procedures, no dynamic control of the voice quality. The approach that we follow here is to apply signal

processing algorithms to the diphones to be concatenated, in the attempt to qualitatively reproduce the differences in the acoustic cues selected to model voice quality.

2.4 Processing of voice units

A graphic tool, displayed in Fig. 1, was developed in Matlab for the off-line processing of short phonetic units (diphones). The tool is intended as a prototyping utility that allows to test new signal processing algorithms on each single diphone in a voice database. Some of the processing functions provided are: spectral emphasis/de-emphasis for spectral-slope control, spectral warping, aspiration noise modelling. No support was provided for pitch related cues, such as jitter, F0 modulations, etc., since diphone concatenation synthesizers relies on pitch control algorithms such as OLA-based processing routines. To date, the tool is compatible with MBROLA diphone databases.

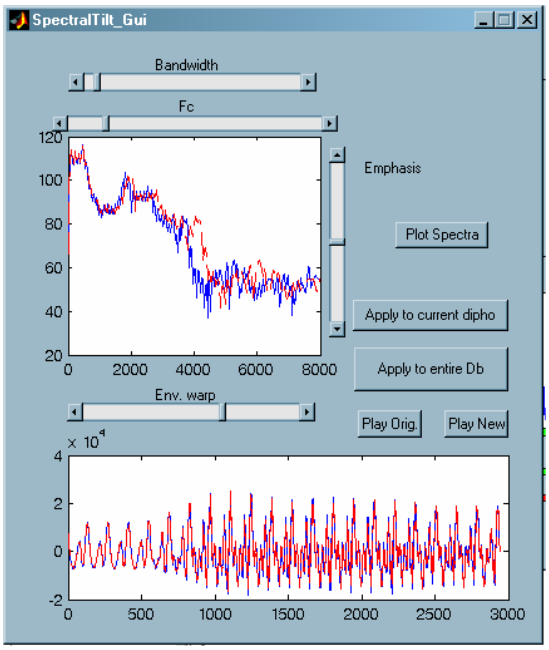

Figure 1: An interactive tool for the design of signal processing of diphones

2.5 Extensions to the Mbrola synthesizer

The diphone processing approach to voice quality control has been implemented by embedding the effects into the synthesizer adopted by us for the Italian speech synthesis, namely the MBROLA diphone concatenation synthesizer. We faced this task by allowing the online processing of the diphones as an intermediate step of the concatenation procedure (see Fig. 2). This step has been implemented using both spectral processing based on DFT and Inverse-DFT transforms, and time-domain processing for pitch-related effects.

The MBROLA speech synthesizer, which originally provides controls for pitch and phoneme duration, has been further extended to allow for control of a set of low-level acoustic parameters that can be combined to produce the desired voice quality effects. Time

evolution of the parameters can be controlled over the single phoneme by means of control curves. The extended set includes gain ("*Vol*"), spectral tilt ("*SpTilt*"), shimmer ("*Shim*"), jitter ("*Jit*"), aspiration noise ("*AspN*"), F0 flutter ("*F0Flut*"), amplitude flutter ("*AmpFlut*"), spectral warping ("*SpWarp*"). A study on how these low-level effects combine to obtain the principal non-modal phonation types encountered in emotive speech is in progress, and more details are reported in a following section on Mark-up language extensions. Here we give a rough description on how these low-level acoustic controls were implemented:
- *Gain ("Vol"):* gain control is obtained by simple rescaling of the spectrum modulus.
- *Spectral tilt ("SpTilt"):*the spectral balance is changed by a reshaping function in the frequency-domain that enhances or attenuates the low- and mid- frequency regions, thus changing the overall spectral tilt.
- *Shimmer ("Shim")*: this is the difference between the amplitudes of consecutive periods. It is reproduced by introducing random amplitude modulations to each consecutive periods of the voiced part of phonemes.
- *Jitter ("Jit")*: this is the period length difference between consecutive periods. It is reproduced by summing random pitch deviations to the pitch control curves computed by Mbrola's prosody matching module.
- *Aspiration noise ("AspN"):* for voiced frames, aspiration noise is generated from the frame DFT transform, by inverse transformation of a high-pass filtered version of the spectral magnitude, and of a random spectrum phase.
- *F0 flutter ("F0Flut"):* random low frequency fluctuations of the pitch are reproduced as for Jitter. The low frequency fluctuations are obtained by random noise band-pass filtering. The second order band-pass filter is tuned in the (4Hz-10Hz) range.
- *Amplitude flutter ("AmpFlut")*: random low amplitude fluctuations are obtained as for Shimmer. The low frequency fluctuations are obtained by random noise band-pass filtering. The second order band-pass filter is tuned in the (4Hz-10Hz) range.
- *Spectral warping ("SpWarp"):* the rising or lowering of upper formants is obtained by warping the frequency axis of the spectrum (through a bilinear transformation), and by interpolation of the resulting spectrum magnitude with respect to the DFT frequency bins.

The Mbrola parser has been modified in order to permit the use of the low-level acoustic controls as general commands or as curves specified at the phoneme level (see the example of an extended phonetic file in Fig. 3).

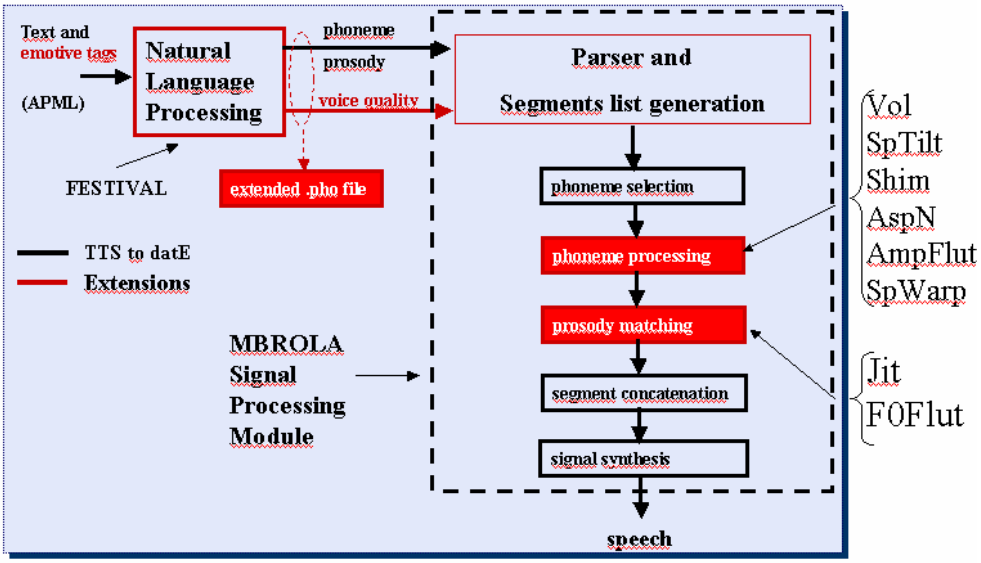

Figure 2: Extensions to the voice synthesis engine (the Mbrola diphone concatenation synthesizer).

```
;Vol=0
;SpTilt=0.0
;Shim=0.0
;Jit=0.0
;AspN=0.0
;F0Flut=0.0
;AmpFlut=0.0
;;SpWarp=0.3

 _ 25 100 143
 a1 309 5 151 20 142 40 150 60 141 80 126 100 116 Shim 0 0.1 100 0.2
 v 85.3333 0 112 50 118 100 127 Shim 0 0.3 100 0.2
 a 334 0 127 20 126 40 118.1250 60 113 80 106 100 148 Vol 0 -3 100 -5 Shim 0 0.2 100
0.4 Jit 0 0.06 100 0.06
 _ 10
```

Figure 3: ***Example of an extended .pho file. The spectral warping command affects all phonemes with constant value 0.3, whereas different gain, shimmer and jitter control curves are specified for different phonemes.***

4. MARK-UP LANGUAGE EXTENSIONS FOR EMOTIVE VOICE SYNTHESIS

The APML markup language for behavior specification allows to specify how to mark up the verbal part of a dialog so as to add to it the "meanings" that the graphical and the speech generation components of an animated agent need to produce the required expressions (De Carolis *et alii*, 2004). So far, the language defines the components that may be useful to drive a face animation through the facial animation parameters (FAP) and facial display functions. A scheme for the extension of a previously developed affective

presentation mark-up language (APML) has been studied (Marchetto, 2004). The extension of such language is intended to support voice specific controls. An extended version of the APML language has been included in the FESTIVAL speech synthesis environment, allowing the automatic generation of the extended .pho file from an APML tagged text with emotive tags. This module implements a three-level hierarchy in which the affective high level attributes (e.g. <anger>, <joy>, <fear>, etc.) are described in terms of medium-level voice qua sed>, etc.). These me c attributes defining t >, <jitter>, etc.). The ıe extended Mbrola sy d above. In Fig. 4, an edium- and low- level

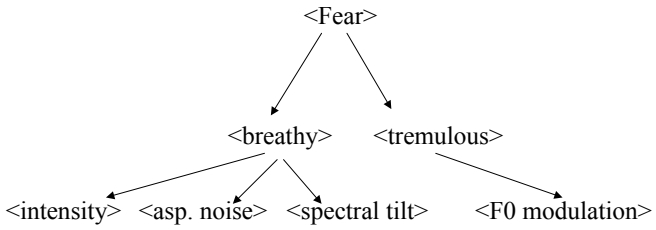

Figure 4: *Qualitative description of voice quality for "fear" in terms of acoustic features*

s betv f env This ı synt f audi ı prog

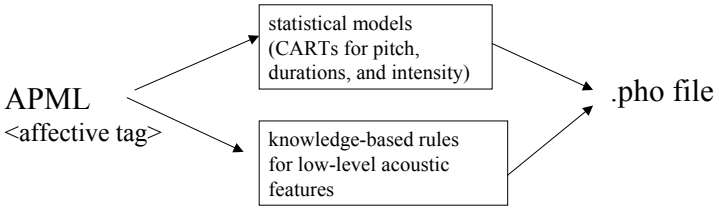

Figure 5: *Detailed representation of the "language processing" block implemented through the APML extensions and the statistical CARTs for prosody.*

Given the hierarchical structure of the acoustic description of emotive voice, we performed preliminary experiments focused on the definition of speaker-independent rules to control voice quality within a text-to-speech synthesizer. Different sets of rules describing the high

and medium level attributes in terms of low-level acoustic cues where used to generate the phonetic files to drive the extended MBROLA synthesizer. Table 1 shows the low level components used to describe the given set of medium level descriptors *soft, loud, whispery, tremulous, hoarse*. The Table reports the control parameter and the activations level of each parameter. Values are in the range [0,1], and have different meanings for the different parameters. E.g., SpTilt=0 means maximal de-emphasis of higher frequency range, whereas SpTilt=0 means maximal emphasis; AspNoise=0 means absence of noise component, whereas AspNoise=1 means absence of voiced component, thus letting aspiration noise component alone; for F0Flut, Shimmer, and Jitter, value=0 means effect is off, whereas value=1 means effect is maximal; SpWarp=0 means maximal spectrum shrinking, and SpWarp=1 means maximal spectrum stretching. Fig. 6 shows the comparison of an example of synthesis obtained with tremulous voice with respect to a similar sentence obtained with modal voice. The tagged text used to generate the synthesis was as follows:

```
<vsml>
<performative type="inform">
<voqual type="modal" level="1.0">Questa e' la mia voce
modale.</voqual><voqual type="tremulous" level="1.0">Questa e' la mia voce
tremante.</voqual>
</performative>
</vsml>
.
```

Table 1: Medium level voice quality description in terms of low-level acoustic components.

	soft	loud	whispery	breathy	tremulous	hoarse
Low level Components	(SpTilt, 0.3)	(SpTilt, 0.7)	(AspNoise, 1.0)	(AspNoise, 0.2) (SpTilt, 0.05)	(F0Flut, 0.9) (SpWarp, 0.3)	(Jitter, 0.3) (Shimmer, 0.1) (AspNoise, 0.2)

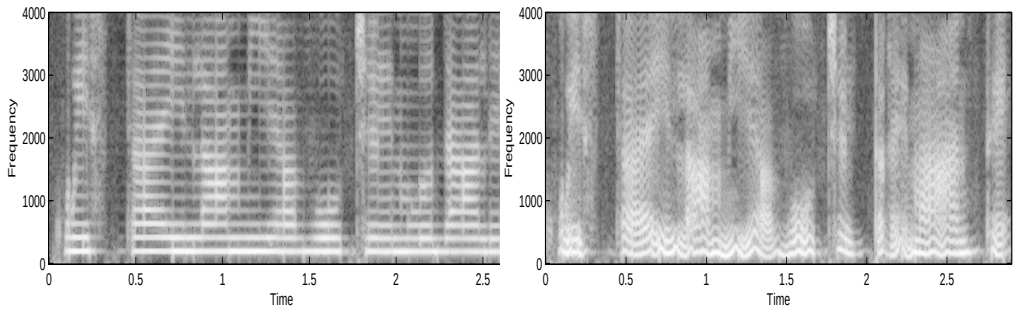

Figure 6: *Spectrograms of the utterance "Questa è la mia voce modale" ("This is my modal voice") in the left panel, and of the utterance "Questa è la mia voce tremante"("This is my tremulous voice") in the right panel. Both utterances were obtained by the modified Festival/MBROLA TTS system using a VSML input text.*

5. AUDIOVISUAL INTEGRATION

Finally, the FAP stream generation components and the audio synthesis components have been integrated into a unique system able to produce the facial animation including emotive audio and video cues, from tagged text. The facial animation framework relies on previous studies for the realization of Italian talking heads (Cosi *et alii*, 2003, Magno Caldognetto *et alii*, 2004). A schematic view of the whole system is shown in Fig. 7. The modules used to produce the FAP control stream (AVENGINE), and the speech synthesis phonetic control stream (FESTIVAL), are synchronized through the phoneme duration information. The output control streams are in turn used to drive the audio and video rendering engines (i.e., the MBROLA speech synthesizer and the face model player).

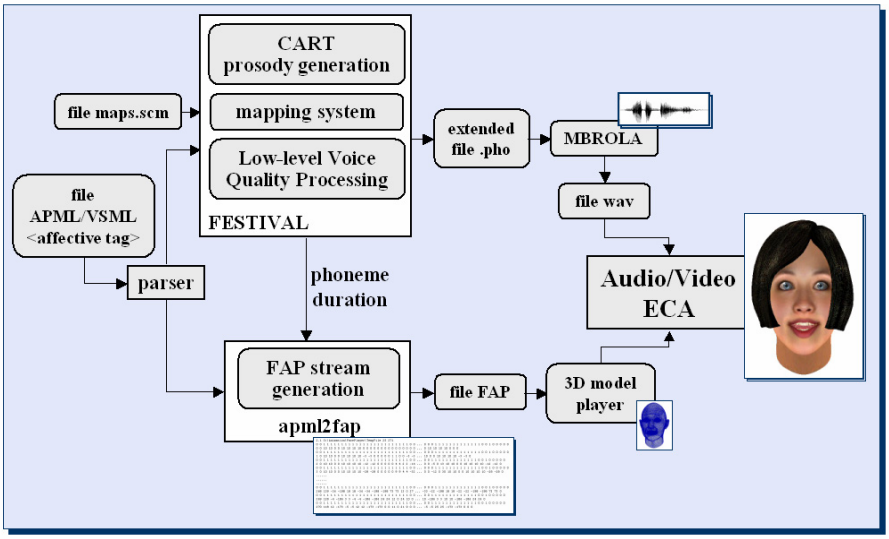

Figure 7: *Block scheme of the system designed to produce the facial animation with emotive audio and video cues, from tagged text.*

6. CONCLUSIONS

The analysis and synthesis of acoustic parameters related to voice quality in emotive speech has been addressed. Speech signal processing techniques have been explored with speaker-independent voice control perspectives. These have been included in the MBROLA diphone concatenation acoustic back-end. The principal limitations of this approach are due to the difficulty in the design of effective speaker-independent signal transformations that should characterize the emotive vocal gestures. To this respect, possible future research directions are foreseen for the study of improved speaker-independent acoustic processing, both from the methodological and from the implementation point of view. Moreover, the use of data-driven techniques to learn the voice transformations from data has to be

evaluated: this could lead to more sophisticated and effective voice transformations, and to more realistic context dependent time-varying control of the transformations.
Finally, the integration between spoken and facial cues has been addressed by connecting a face control engine and the speech synthesis engine used for expressive synthesis from tagged text. The resulting system allows to generate a speech synthesis control stream (PHO file) and a facial animation control stream (FAP file), which are used in turn to produce synchronized and emotionally coherent audio and video outputs.

AUDIO EXAMPLES

- from soft to loud: spectral tilt processing on vowel /a/ (male voice) [play].
- formant shift: spectral warping on diphone /dZ-a1/ (male voice) [play warp down, play warp up].
- sequence of seven sentences synthesized with different voice qualities (female voice): 1. modal ("Questa è la mia voce modale"), 2. soft ("Questa è la mia voce bassa e morbida"), 3. loud ("Questa è la mia voce forte"), 4. whispery ("Questa è la mia voce sussurrata"), 5. breathy ("Questa è la mia voce aspirata"), 6. tremulous ("Questa è la mia voce tremante"), 7. hoarse ("Questa è la mia voce roca") [play].

ACKNOWLEDGEMENTS

Part of this work has been sponsored by, PF-STAR (Preparing Future multiSensorial inTerAction Research, European Project IST- 2001-37599, http://pfstar.itc.it/).

BIBLIOGRAPHY

Cosi P., Fusaro A., Tisato G. (2003), LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model, in *Proceedings of Eurospeech 2003*, Geneva, Switzerland, September 1-4, 127-132.

d'Alessandro C., Doval B. (1998), Experiments in voice quality modification of natural speech signals: the spectral approach, in *Proceedings of the 3rd ESCA/COCOSDA Int. Workshop on Speech Synthesis*, 277–282.

De Carolis B., Pelachaud C., Poggi I., Steedman M. (2004), APML, a Markup language for believable behavior generation, in Book *Life-Like Characters, Tools, Affective functions, and Applications,* H. Prendinger and M. Ishizuka Eds., Springer.

Drioli C., Avanzini F. (2003), Non-modal voice synthesis by low-dimensional physical models, in *Proc. of the 3rd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, Florence, Italy, December 10-12.

Drioli C., Tisato G., Cosi P., Tesser F. (2003), Emotions and voice quality: experiments with sinusoidal modeling, *in Proc. of Voice Quality: Functions Analysis and Synthesis (VOQUAL) Workshop*, Geneva, Switzerland, August 27-29, 127-132.

Gobl C., Chasaide A.N. (2003), The role of the voice quality in communicating emotions, mood and attitude, *Speech Communication*, vol. 40, 189–212.

Johnstone T.,. Scherer, K.R. (1999), The effects of emotions on voice quality, *in Proceedings of the XIV Int. Congress of Phonetic Sciences*, 2029–2032.

Ladd D. R., Silverman K.E.A., Tolkmitt F., Bergmann G., Scherer K.R. (1985), Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect, *Journal of the Acoustical Society of America*, vol. 78, n. 2, 435–444.

Magno Caldognetto E., Cosi P., Drioli C., Tisato G., Cavicchio F. (2004), Modifications of phonetic labial targets in emotive speech: effects of the co-production of speech and emotions, *Speech Communication*, vol. 44, n. 1-4 , 173-185.

Marchetto E. (2004), *Sistema per il controllo della voice quality nella sintesi del parlato emotivo*, MThesis, Univ. of Padova, Italy.

Schröder M., Grice M. (2003), Expressing vocal effort in concatenative speech, in *Proceedings of 15th ICPhS*, Barcelona, Spain, 2589–2592.

Tesser F., Cosi P., Drioli C., Tisato G. (2004), Prosodic data driven modelling of a narrative style in FESTIVAL TTS, *in Proc. of the 5th ISCA Speech Synthesis Workshop*, Pittsburgh, USA, June 14-16, 185-190.