eNTERFACE'05
july 18- August 12, 2005 – Faculté Polytechnique de Mons,- Belgium

SPEECH CONDUCTOR

Project Title:
Speech Conductor

Principal investigator:
Christophe d'Alessandro (LIMSI-CNRS)

Candidates:
Sylvain Le Beux (MASTER SETI, Paris XI university)

Abstract:

The Speech Conductor project aims at developing a gesture interface for driving ("conducting") a text to speech synthesis system. Then, automatic speech synthesis will be modified in real time according to the gestures of a "Speech Conductor". The Speech Conductor will add expression and emotion to the speech flow using speech signal modification algorithms and gesture interpretation algorithms.

PROJECT OBJECTIVE

Speech synthesis quality seems nowadays acceptable for applications like text reading or information playback. However, these reading machine lack expression. This is not only a matter of corpus size, computer memory or computer speed. A speech synthesizer using several times more resources than currently available will probably improve on some points (less discontinuities, more smoothness, better sound), but will not be able of more expression. Fundamental question concerning expression in speech are still unanswered, and to some point even not stated. Expressive speech synthesis is the next challenge.

Expressive speech synthesis may be viewed on two sides: expression specification (what expression in this particular situation?) and expression realisation (how the specified expression is actually implemented). The first problem (situation analysis and expression specification) is one of the most difficult problems for computational linguistics research: understanding a text and its context. Without deep knowledge of the situation expression is nonsense.

It is only the second problem that will be addressed in this workshop. Given the expression specification, let say the "expression score" for a given text, how to "interpret" it according to this score?

The Speech Conductor project aims at developing and testing gesture interfaces for driving ("conducting") a text to speech synthesis system. Then, it will be possible to modify speech synthesis in real time according to the gestures of the "Speech Conductor". The Speech Conductor will add expression and emotion to the speech flow using speech signal modification algorithms and gesture interpretation algorithms. This will be one step towards expressive speech synthesis

This project has a strong applicative side (e.g. augmented expressive speech capabilities for disabled people etc) and a strong research side (rules for controlling expression, algorithms for speech quality modifications and gesture capture).

Evaluation of the results will take place at an early stage in the design and development process. As no specific evaluation methods for expressive speech are currently available, it will be necessary to spend some time for designing methods or at least conduct a reflection on that matter. Ultimately expressive speech could be evaluated through a modified Turing test or behavioural testing.

This project is fundamentally multimodal. Output of the system involves the auditory modality (and possibly latter in the project the visual modality through using an animated agent). Input modalities are text, gestures, and possibly vision.

The project objectives during the workshop are to conduct a research, to propose solutions and to implement prototypes for the following problems:

1. speech parameter control for expressive synthesis
2. speech signal parametric modification
3. gestures capture (may be including video)
4. gestures to parameter mapping
5. gesture speech controller architecture
6. prototype implementation using a Text to Speech system and/or a parametric synthesiser
7. expressive speech assessment methodologies

TECHNICAL DESCRIPTION

A Technical description:

The main goal of this project is to test various gesture interfaces for driving a speech synthesiser. It is then hoped that expressive speech more “natural” than rule-based or corpus-based expressive speech will be produced. Of course this project is too much ambitious for 4 weeks of work, but we think that significant insights into this project could be gained in the framework of eNTERFACE’05. The main technical points to be addressed are:

1. Identify the parameter of expressive speech and their relative importance. All the speech parameters are supposed to vary in expressive speech. A list of signal domain parameters would encompass: articulation parameters (speed of articulation, formant trajectories, articulation loci, noise bursts, etc.) phonation parameters (fundamental frequency, durations, amplitude of voicing, glottal source parameters, degree of voicing and source noise etc.). Alternatively, physical parameters (sub glottal pressure, larynx tension) would be used.
2. Signal processing for expressive speech. Techniques for parametric modification of speech: fundamental frequency, durations, articulation rate, voice source.
3. Domain of variation and typical patterns for expressive speech parameters, analysis of expressive speech. To some point, it will be necessary to analyse real expressive speech for finding patterns of variation.
4. Gestures capture and sensors. Many types of sensor and gesture interfaces are available. The most appropriates would be selected and tried.
5. Mapping between gestures and speech parameters. The correspondence between gestures and parametric modifications is of paramount importance. This correspondence can be more or less complex (one to many, many to one, one to one).
6. Different types of speech synthesis could be used. Ideally, the most appropriate would be fully parametric physical synthesis. However, this is neither the most efficient nor the most widespread type of TTS system. Then diphone base concatenative speech synthesis or formant synthesis could be also used. Real time implementations of the TTS system are needed. A part of the work will be devoted to these implementations.
7. Expression, emotion, attitude, phonostylistics. When the system will be ready; selected questions and hypotheses in the domain of emotion research and phonostylistic will be revisited.
8. Evaluation methodology for expressive speech synthesis will be addressed. Preliminary evaluation of the results obtained will take place at an early stage of the project.

B. Resources needed: facility, equipment, software, staff, etc.

Facility:

The project involves speech synthesis and interfaces. A quiet room would be desirable, because at some early point in the project, sound will be produced: then it will be important to avoid annoying other people working on different project. Some space for interfaces would also be necessary: e.g. tables for one or more digital master keyboards, joysticks

Equipment:

Computers: personal computers (PC or Mac) with quality output sound device.

Interfaces: MIDI device (for PC, built in for Mac). MIDI keyboard, joysticks, and corresponding hardware drives on the computer
Audio: quality loudspeakers and headphones.

Software:
Speech synthesis system(s), e.g. Mbrola.
Speech analysis systems e.g. Praat
Musical programming environment, e.g. Max
General purpose programming environment: C compiler, …

Staff: system administrator for preparing the computers, giving access to the network etc.

C. Project management

Pre-workshop

The first task is selection of the team members and constitution of the team. As soon as the team is defined, a discussion list will be established. Discussion on the project prior to the workshop itself will be useful for presenting the team members, for preparing the workshop in terms of software and hardware and for scientific discussion.

Workshop

As the workshop involves a small team and close cooperation, management will be reduced to a minimum. The main tasks of management will be to organize the work, propose the schedule and follow production of work and document. Meetings will be organized as needed, at least the first and last half-day of each week.

Post-workshop

The discussion list will last after the meeting. Continuation of the project (e.g. publications, collaboration) will be envisaged if appropriate.

EXPECTED OUTCOME OF THE PROJECT

A working prototype for controlling a speech synthesiser using a gesture interface should be produced at the en of the project. Another important outcome is the final report which will contain a description of the work and the solved and unsolved problems. This report could serve as a basis for future research in the domain and a conference or journal publication.

WORKPLAN AND IMPLEMENTATION SCHEDULE

Four full weeks are allotted to the project. A minimum preliminary work can be done by advance (e.g. papers reading, literature survey, team discussions). Each week will end and start with a team meeting and report to other eNTERFACE'05 projects for general discussion and exchanges.

As for computer programming the main tasks are: 1. to implement real-time versions of synthesis systems. 2. to map gesture control output parameters on synthesis input parameters.3. to implement gesture controlled parametric speech modifications.

1. Week 1

   In the first week, the main goal is to define the system architecture, and to assemble the hardware and software pieces that are necessary. Some time is also devoted to evaluation methodology and general discussion and exchanges on expressive speech and synthesis. At the end of the first week, the building blocks of the system (i.e. TTS system, gesture devices …) should be running separately. The system architecture and communication protocols should be defined and documented.

   a. Day 1: opening day, first week opening meeting, tutorial 1
   b. Day 2 discussion, system design and implementation
   c. Day 3 (Belgium national day) discussion, system design and implementation
   d. Day 4 discussion, system design and implementation
   e. Day 5 discussion, system design and implementation. First week closing meeting, work progress report 1: architecture design, final work plan

2. Week 2

   The main work in the second week will be implementation and test of the gesture based speech control system. At the end of the second week, a first implementation of the system should be near to ready. This includes real time implementation of synthesis software and fusion between gesture and synthesis control parameters.

   a. Day 1 2nd week opening meeting, tutorial 2. System implementation and test.
   b. Day 2 system implementation and test.
   c. Day 3 system implementation and test.
   d. Day 4 system implementation and test.
   e. Day 5 system implementation and test.2nd week closing meeting, work progress report 2

3. Week 3

   The main work in the third week will be implementation and test of the gesture based speech control system. At the end of the third week, an implementation of the system should be ready. Expressive speech synthesis patterns should be tried using the system.

   a. Day 1 3rd week opening meeting, tutorial 3. System implementation, expressive synthesis experiments.
   b. Day 2 System implementation, expressive synthesis experiments.
   c. Day 3 System implementation, expressive synthesis experiments.
   d. Day 4 System implementation, expressive synthesis experiments.
   e. Day 5 3rd week closing meeting, work progress report 3. System implementation, expressive synthesis experiments.

4. Week 4

   The 4th week is the last of the project. Final report writing and final evaluation are important tasks of this week. The results obtained will be summarized and future work will be envisaged for the continuation of the project. Each participant will write an individual evaluation report of the project in order to assess its success and to improve organisation and content of future similar projects.

a. Day 1 4th week opening meeting, tutorial 4
b. Day 2 implementation, evaluation, report,.
c. Day 3 implementation, evaluation, report, demonstration preparation.
d. Day 4 implementation, evaluation, report, demonstration preparation.
e. Day 5 closing day, final meeting, final report, demonstration, evaluation

PROFILE OF TEAM

A. Leader

Christophe d’Alessandro.
Christophe d'Alessandro was born in Marseille, France, on December 16, 1961. He received the B.S. degree in Mathematics, the M.S and the Ph.D degrees in Computer Science from Paris VI University, in 1983, 1984 and 1989, respectively. He has been a permanent Researcher at LIMSI, a laboratory of the CNRS (French National Agency for Scientific Research), since october 1989. Prior to joining the CNRS, Dr. d'Alessandro has been a Lecturer in computer science at Paris XI University from october 1987 to october 1989. He also graduated in music, and he has been appointed in 1988 organist of the historical organ of Sainte-Elisabeth in Paris, France (titular organist since 1992).
His research interests include text-to-speech synthesis, signal processing for speech analysis and synthesis, perception and synthesis of intonation in speech and singing, voice source analysis and synthesis, speech synthesis assessment, musical acoustics, musicology. He is a member of ASA, ESCA, IEEE, ATALA and SFA (French Acoustical Society). At LIMSI, he is the head of the Situated Perception Group since 2003. He is also involved in researches on historical instruments and music, and he is a member of the historical monument committee at the French ministry of culture (historical instruments).

B. Staff

To be defined.

C. Other researchers needed

This project is by nature multidisciplinary. Ideally, researchers with strong backgrounds in one or more of the following topics would be needed: gesture processing, multimodal interfaces design, psychology of emotion, speech processing, phonostylistics, and any other relevant domain.

REFERENCES

Interfaces and gesture

M. Wanderley and P. Depalle, “Gestural Control of Sound Synthesis”, Proc. of the IEEE, 92, 2004, p. 632-644.
“MIDI musical instrument digital interface specification 1.0,” Int. MIDI Assoc., North Hollywood, CA, 1983.
S. Fels, “Glove talk II: Mapping hand gestures to speech using neural networks,” Ph.D. dissertation, Univ. Toronto, Toronto, ON, Canada, 1994.

Text to speech

Dutoit T. An Introduction to Text-To-Speech Synthesis. Kluwer Academic Publishers, 1997.
Klatt D., Review of text-to-speech conversion for English, (with a LP record) J. Acoust. Soc. Am., Vol. 82, 737-793. 1987.
C. d'Alessandro. « 33 ans de synthèse de la parole à partir du texte: une promenade sonore (1968-2001) ». Traitement Automatique des Langues (TAL), Hermès, Vol. 42 No 1, p. 297-321, (with a CD 62 mn), 2001 (in French)

Emotion, speech, Voice quality

C. d'Alessandro, B. Doval, "Voice quality modification for emotional speech synthesis", Proc. of Eurospeech 2003, Genève, Suisse, pp. 1653-1656
M. Schröder “Speech and emotion research”, Phonus, Nr 7, june 2004 ISSN 0949-1791, Saarbrücken
Various authors: Speech Communication. Special issue Speech and Emotion, 40(1-2), 2003.

FORESSEN INVITED SPEAKERS

This is a tentative list of proposed/foreseen invited speakers and/or interesting topics in the framework of the workshop. This list is submitted to the workshop organiser, more proposals would be made if needed.

Multimodal interfaces :
Jean-Claude Martin (LIMSI, also in the HUMAINE Network)

Musical programming environments:
Somebody at Ircam ? Norbert Schnell, François Déchelle

Gesture interfaces and musical expression:
Marcello Wanderley, Philippe Depalle

Text to speech synthesis.
Thierry Dutoit, C. d'Alessandro

Speech and emotion, prosody and emotion
Marc Schröder, Piet Mertens, C. d'Alessandro

Gestures and speech

(Note: the team leader (C. d'Alessandro) will be absent on Tuesday july 28, Friday july 29, and Monday august 1.)