Towards Expressive Speech Synthesis in English on a Robotic Platform

Sigrid Roehling, Bruce MacDonald, Catherine Watson

Department of Electrical and Computer Engineering
University of Auckland, New Zealand
s.roehling, b.macdonald, c.watson@auckland.ac.nz

Abstract

Affect influences speech, not only in the words we choose, but in the way we say them. This paper reviews the research on vocal correlates in the expression of affect and examines the ability of currently available major text-to-speech (TTS) systems to synthesize expressive speech for an emotional robot guide. Speech features discussed include pitch, duration, loudness, spectral structure, and voice quality. TTS systems are examined as to their ability to control the features needed for synthesizing expressive speech: pitch, duration, loudness, and voice quality. The OpenMARY system is recommended since it provides the highest amount of control over speech production as well as the ability to work with a sophisticated intonation model. OpenMARY is being actively developed, is supported on our current Linux platform, and provides timing information for talking heads such as our current robot face.

1. Introduction

Affect influences speech, not only in the words we choose, but in the way we say them. These vocal nonverbal cues are important in human speech as they communicate information about the speaker's state or attitude more efficiently than the verbal content (Eide, Aaron, Bakis, Hamza, Picheny, and Pitrelli 2004). We wish to include vocal cues in synthetic speech in order to allow more efficient and more pleasant communication between humans and robots. Effective and socially appropriate human-robot communication becomes more and more important as robots are increasingly used in social situations (Breazeal 2004), for example as guide robots. Several researchers have attempted to isolate aspects of speech that are related to a particular emotion or a particular emotion dimension. The research methods, emotions, and speech parameters that were previously considered vary greatly and accordingly the results have been varied, sometimes even contradictory. In Section 2 we review the results available and attempt to categorize them in order to determine which aspects of speech are important in conveying affective information. Section 3 discusses various text-to-speech systems available, both free and commercial, and assesses their capacity to produce expressive speech. A brief description of our expressive robot project can be found in Section 4. Section 5 evaluates the applicability of these TTS systems for a robot platform such as ours. Section 6 concludes.

2. Importance of Vocal Features in the Expression of Affect

Beyond making the verbal content highly intelligible, expressive speech research is based on the importance of vocal nonverbal cues in communicating affect as shown by higher rates of recognition of emotions from recorded speech than from the transcribed text alone (Scherer, Ladd, and Silverman 1984). The research has focused on finding the specific features of speech that convey emotional information. These features are examined below. Where not explicitly stated otherwise the research is concerned with the English language.

2.1. Pitch

Pitch contour seems to be one of the clearest indicators of affect according to Williams and Stevens (1972); it was found to be relevant in German (Burkhardt and Sendlmeier 2000) and Dutch (Mozziconacci and Hermes 1999), but Frick (1985) and Gobl, Bennett, and Chasaide (2002) (Swedish) contend that pitch alone is not sufficient. In particular, the interaction of pitch with loudness (Frick 1985) and with the grammatical features of the text (Scherer, Ladd, and Silverman 1984) (German) seems to be critical. When synthesizing more subtle emotions, Burkhardt and Sendlmeier (2000) found intonation pattern to be relevant in German but even including vowel precision and phonation type it was not sufficient to distinguish between crying despair and fear as well as between quiet sorrow and boredom. Cahn (1990) in her pioneering experiments determined that emotion mostly affects correlates related to pitch and duration, and Vroomen, Collier, and Mozziconacci (1993) found that in Dutch pitch and duration are sufficient to distinguish between neutral speech, joy, boredom, anger, sadness, fear, and indignation. In contrast, pitch has been found to carry little information by Edgington (1997) for English and Heuft, Portele, and Rauth (1996) for German.

2.2. Duration

Duration is one of the correlates affected the most by emotion according to Cahn (1990) and together with pitch it is sufficient to distinguish between neutral speech, joy, boredom, anger, sadness, fear, and indignation in Dutch (Vroomen, Collier, and Mozziconacci 1993). Edgington (1997) and Heuft, Portele, and Rauth (1996), however, claim that duration carries little emotional information in English and German, respectively.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

2.3. Loudness

Findings mentioned in Frick (1985) indicate that loudness may not be important but he admits that correct synthesis of loudness may simply be difficult.

2.4. Spectral Structure

Scherer, Ladd, and Silverman (1984) claim that spectral energy distribution (together with voice quality) carries much of the affective information in German. Both Frick (1985) and Lee, Yildirim, Bulut, Kazemzadeh, Busso, Deng, Lee, and Narayanan (2004) confirm that spectral structure is significant. Rank and Pirker (1998) noted that adding spectral energy distribution (as well as voice quality, harmonics-to-noise ratio, and articulatory precision) increased recognition rates in Austrian German. However, energy has been found to carry little information by Edgington (1997) for English and Heuft, Portele, and Rauth (1996) for German.

2.5. Voice Quality

Scherer, Ladd, and Silverman (1984) claim that voice quality is one of the aspects that carry much of the affective information in German. Voice quality is significant in German (Burkhardt and Sendlmeier 2000) and Swedish (Gobl, Bennett, and Chasaide 2002). Rank and Pirker (1998) noted that adding voice quality increased recognition rates in Austrian German. When synthesizing more subtle emotions in German, Burkhardt and Sendlmeier (2000) found vowel precision and phonation type relevant but (even with intonation pattern) not sufficient to distinguish between crying despair and fear as well as between quiet sorrow and boredom.

2.6. Suprasegmental Versus Segmental Elements

Speech parameters can be categorized into segmental and suprasegmental (prosodic) features. This is most clearly observed in concatenative synthesis which combines prerecorded human speech segments to form utterances. Segmental features are those that pertain to a single speech segment and in concatenative synthesis are controlled for by selecting segments from a specific prerecorded corpus. Suprasegmental features, such as pitch contour, pertain to the utterance as a whole and in concatenative synthesis are usually controlled for by applying signal processing techniques to the chosen sequence of speech segments. Montero, Gutiérrez-Arriola, Colás, Enríquez, and Pardo (1999) discuss experimental results which suggest that in Spanish some emotions are expressed more clearly through segmental features whereas the expression of other emotions focuses on suprasegmental elements. This was confirmed for English by Bulut, Narayanan, and Syrdal (2002) whose experiments showed that anger is predominantly expressed through segmental features while neutral speech and sadness are prosody dominant. Highest recognition rates are achieved when both segmental and suprasegmental features are related to the same emotion.

2.7. Summary

While there is some disagreement as to how much each of the aspects mentioned contributes to expressing emotion, none of those investigated here can be clearly ruled out as not being relevant. Since we are developing a research platform for investigating expressive speech we want to be able to control all of these parameters. The specific features we focus on in our review of TTS systems are therefore ability to control parameters related to pitch, duration, loudness, spectral structure, and voice quality.

3. Overview of Available Text-to-Speech Systems That Can Potentially Create Expressive Speech

A number of TTS systems were investigated. The selection presented here constitutes those that were advertised as being able to synthesize expressive speech or were otherwise deemed to have potential in this area because of the number of controllable speech parameters. The features of the TTS systems are summarized in Table 1 and discussed in detail below, with regards to the required features from the previous section. None of these systems specifically address spectral structure, thus this parameter has been omitted from the table. Pitch, as well as playing a role in conveying emotion, also has a semantic function. The phrase "he went home", for example, is interpreted as a statement if spoken with falling intonation and as a question if spoken with a rising pitch on "home". Because correct intonation is crucial to understanding meaning, listening to synthetic speech with poor intonation is tiring. In the interest of creating expressive synthetic speech that is pleasant to listen to we also examine the TTS systems as to whether they allow us to use a sophisticated intonation model.

3.1. Commercial Systems

3.1.1. Acapela BabTTS

BabTTS comes with two different types of voices. High Density (HD) voices are diphone MBROLA voices whereas High Quality (HQ) voices are based on unit selection. The system supports a C/C++ API, SAPI, and the proprietary NSCAPI (Acapela Group 2005b). Via SAPI XML tags one can control baseline pitch for HD voices, pauses, speech rate, and volume (Acapela Group 2005a). It runs on Windows, Linux, and Mac OS (A. Bistes, personal communication, October 1, 2006).

3.1.2. AT&T Natural Voices

Previously known as the research project Next-Gen, AT&T views this system as the successor to the Bell Labs TTS system (H. J. Schroeter, personal communication, July 28, 2006). It employs unit selection and can control overall speech rate and overall volume. Through markup speech rate and volume can be controlled at the word level. Silence fragments for which duration can be specified can also be passed. A C++ API and SDK are provided and it supports SAPI, SSML, and JSML. Windows, Linux, and Solaris are the major platforms supported (AT&T 2002).

3.1.3. Fonix DECtalk 4.6.4

Fonix DECtalk is the successor to Digital's DECtalk. It is based on formant synthesis and is therefore able to control a large number of speech parameters (Hallahan 1995). Pauses, speech rate, volume, and voice quality parameters

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

TTS SYSTEMS	SPEECH PARAMETERS				
	Pitch	**Duration**	**Loudness**	**Voice Quality**	**Notes**
BabTTS	baseline	pause duration, speech rate	volume		all word-level
Natural Voices		pause duration, speech rate (overall, word-level)	volume (overall, word-level)		
DECtalk	baseline, range, assertiveness, stress rise, accent	pause duration, speech rate	volume	breathiness	all word-level
Naxpres	contour, syllable pitch	phone duration	contour		all controlled automatically by system
Loquendo	baseline	speech rate	volume		all word-level, emphasis
Mulan	pitch	segment duration			all controlled automatically by system
Speech SDK	word-level pitch	overall and word-level speech rate	overall and word-level volume		word emphasis
RealSpeak	sentence accent	pauses, speech rate	volume		all word-level
Festival	baseline, range (both sentence-level), word-level pitch	speech rate (sentence- and word-level)	word-level volume		intonation contour via ToBI
gnuspeech	pitch		amplitude		glottal pulse shape, harmonic content, all not controllable by user
OpenMARY	baseline, range (both sentence-level)	speech rate (sentence-level), pause and segment duration (word-level)	vocal effort (German voice only)	articulation precision	intonation contour via ToBI
ProSynth	system decides pitch and duration based on phrase, accent group, and foot boundaries				

Table 1: Comparison of which speech parameters can be controlled by the various TTS systems.

such as breathiness can be modified dynamically through inline commands embedded in the text to be spoken. Alternatively, word accent, speech rate, pauses, volume, and pitch can be controlled via SAPI XML tags. Because of the use of tags, most of these changes would affect at least one complete word. Operating systems supported include Windows, Mac OS X, and Linux and an SDK is provided (Fonix nd).

3.1.4. IBM Naxpres

IBM's research TTS system is based on concatenation. By recording different corpora for four different speaking styles (questioning, contrastive emphasis, and conveying good and bad news) some expressive speech has been realized through text markup. The underlying system exerts control over pitch contour, syllable pitch, and phone duration, as well as loudness contours, although the latter is not currently used (Pitrelli, Bakis, Eide, Fernandez, Hamza, and Picheny 2006).

3.1.5. Loquendo

Loquendo's TTS system for Windows and Linux uses a unit selection method and a proprietary low-level synthesizer (E. Zovato, personal communication, August 24, 2006). Pitch, speech rate, volume, emphasis, and pauses can be controlled by means of escape sequences such as `\pitch=10 This is a high pitch sentence`. Pitch, rate, volume, emphasis, and pauses can also be controlled through SSML tags like so:
`<prosody rate=-20%> I am tired </prosody>`.
It comes with an SDK and supports SAPI, a C/C++ API, and SSML (Loquendo 2005).

3.1.6. Microsoft Mulan

Mulan is a Chinese-English bilingual system based on unit selection. Its main goal is to provide a system that can process both English and Chinese in the same sentence while preserving sentence level intonation. Pitch and dura-

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

tion are controlled automatically by the system at the unit level. Further work will involve trying to implement different speaking styles (Chu, Peng, Zhao, Niu, and Chang 2003).

3.1.7. Microsoft Speech SDK

This is a concatenative system for Windows. Overall rate and overall volume can be controlled through the API. On a word level, inline tags can be used to control pitch, word emphasis, speech rate, and volume. It supports SAPI (Microsoft 2006).

3.1.8. Nuance RealSpeak

Formerly known as ScanSoft RealSpeak, this TTS system uses a unit selection approach to synthesize speech (F. Broicher, personal communication, July 27, 2006). Volume, speech rate, pauses, and sentence accent can be set through the use of commands in line with the text. These changes are all word based. The major operating system supported is Windows. An SDK and C/C++ or VB API is provided. SAPI support is also included (ScanSoft 2004).

3.2. Non-Commercial Systems

3.2.1. Festival

Festival is a text-to-speech research framework actively developed by the Centre for Speech Technology Research (CSTR) at the University of Edinburgh. It uses a residual excited LPC synthesizer, but also supports MBROLA. Synthesis is diphone based. Pitch baseline, pitch range, and speech rate for the whole sentence can be controlled. Using ToBI one can also control the intonation contour to some extent. For example, the question "Would you like a coffee?" would be marked up as

```
Would you (like ((accent H*))) a
(coffee ((accent H*)(tone L-H%))).
```

Pitch, speech rate, and volume can also be set through SABLE tags embedded in the text, for example,

```
Without his penguin, <PITCH
BASE="-20%"> which he left at home,
</PITCH> he could not enter the
restaurant.
```

However, Festival can only process SABLE markup from files and not on the command line. It runs on Unix and has APIs for several programming languages (Black, Taylor, and Caley 1999).

3.2.2. Gnuspeech

The continuation of a Trillium Sound Research Inc. project, this system is a tube-model-based articulatory synthesizer. In theory many vocal tract parameters, including glottal pulse shape, harmonic content, pitch, and amplitude can be controlled using the tube model, however, this hasn't been implemented in the user interface yet. The system's native platform is NeXT and it is currently being ported to Linux by the GNU project (Hill nd).

3.2.3. OpenMARY

The MARY system was developed by the German Research Center for Artificial Intelligence (DFKI) and Saarland University for German text-to-speech, but now also supports English. It uses the Festival framework and MBROLA as the low level diphone synthesizer. Just as Festival it supports ToBI and can therefore control sentence level speech rate, pitch baseline, and range as well as intonation contour. In addition, pitch accent, pauses, segment duration, intonation contour, and articulation precision can all be controlled by MaryXML tags that are placed within the text. For example, the phrase "Today is a really nice day" (with evaluation of +100, activation of 0, power of +50) would be turned into

```
<prosody pitch="+5%"
pitch-dynamics="-30%" range="4st"
range-dynamics="-20%"
preferred-accent-shape="alternating"
accent-slope="-50%"
accent-prominence="-50%"
preferred-boundary-type="low"
rate="+20%" number-of-pauses="+0%"
pause-duration="-0%"
vowel-duration="+45%"
nasal-duration="+45%"
liquid-duration="+45%"
plosive-duration="-30%"
fricative-duration="-30%" volume="50">
Today is a really nice day.
</prosody>
```

OpenMARY runs on Windows, Linux, MacOS X, and Solaris and supports SSML and APML. Timing information for talking heads can also be provided. The system is maintained by the DFKI and is being actively developed (MARY Text To Speech nd).

3.2.4. ProSynth

This system evolved from a joint project between the University of Cambridge, University College of London, and the University of York from 1997 to 2000. It can drive both MBROLA as a diphone synthesizer and HLsyn as a formant synthesizer. The web demo can also use Festival for synthesis. The user has little control over speech parameters, one can merely mark intonation phrase boundaries, accent group boundaries, and foot boundaries using diacritics embedded in the text. For example,

```
this is a 'sentence / this is an'other
one.
```

ProSynth runs on Windows (ProSynth Project Deliverables nd).

4. Description of Robot Project

The Robotics Lab at the University of Auckland is developing an empathetic robot who will recognize emotion from vocal cues and respond with an appropriate expressive voice and facial expression. We envision this robot to function for example as a guide in a museum. In this capacity the robot need not make use of all emotions, so we will focus on the ones appropriate for this particular situation, e.g. the robot should be able to greet people, sound happy and excited about what it tells them, but also be authoritative when delivering facts and convey a sense of urgency when informing others about its battery running low. Ideally, communication of emotion will happen through three channels, verbal, vocal nonverbal,

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

Figure 1: Modified B21r robotic platform with virtual face.

and facial, each complementing the other two. This project also provides a robotic platform on which to experiment with expressive speech and the interaction between speech and facial expression.

The robot is based on a B21r robotic platform with sonar, infrared, and contact sensors. It includes a stereo color camera system, a laser range finder, and a wireless network link, and is controlled by a dual processor PC running Linux. Hardware abstraction is provided by the player architecture (The Player Project nd; Gerkey, Vaughan, and Howard 2003). For communicating with humans there are speakers, a microphone, and a monitor as its head. The robot (shown in Figure 1) has a 3D virtual face which is capable of expressing several emotions as well as rendering the correct lip movements for speech. This is realized through the use of visemes which are the visual representations of a mouth uttering a sound. Each sound has a corresponding viseme which is displayed at the same time the sound is played. The four sounds and visemes for the word "coffee" are shown in Figure 2. The current system uses Festival as the speech synthesizer. We have added the capability to send ToBI-annotated sentences to the rudimentary Festival driver of the player architecture. For example,

```
say (Would you (like ((accent H*))) a
(coffee ((accent H*)(tone L-H%)))).
```

5. Evaluation of Text-to-Speech Systems for Use in the Robot Project

As emotions have physical effects on, among other things, the vocal tract, an articulatory TTS system such as gnuspeech would suit our purposes best but unfortunately the development in that area has not reached the level of sophistication we are looking for. Among the other systems mentioned here, DECtalk, Festival, and OpenMARY stand out as they allow control over a larger number of parameters related to pitch, duration, and loudness. DECtalk in particular has more control options than could be listed here, however, it does not provide a mechanism for easy markup of intonation contours such as ToBI. Both Festival and OpenMARY support ToBI and OpenMARY can also control articulation precision. We thus recommend OpenMARY for expressive speech synthesis. We currently use Festival with ToBI. OpenMARY is being actively developed, supported on our current Linux platform, and provides timing information for talking heads such as we are using now, therefore we expect a switch to encur few obstacles.

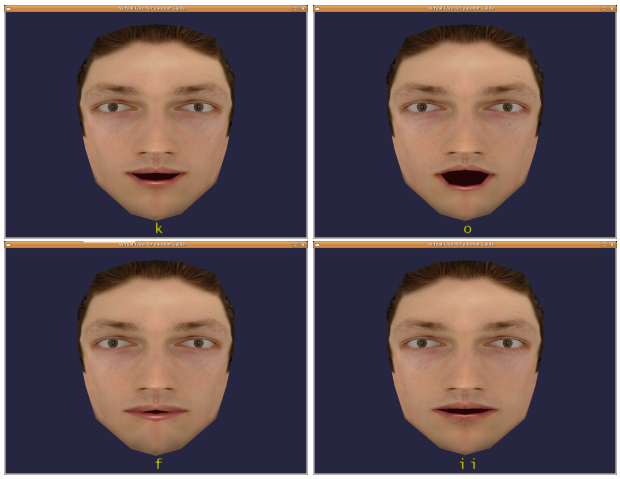

Figure 2: The four phones /k o f ii/ in the word "coffee" and their visemes.

6. Conclusions

We have summarized research on vocal correlates of emotion and categorized the results. We found pitch, duration, loudness, spectral energy structure, and voice quality to be important for synthesizing expressive speech. A review of available TTS system with regards to these findings concluded that OpenMARY provides the best solution for an expressive robot project such as ours.

7. Glossary

API Application Programming Interface.

Low-level synthesizer Takes a phonetic description and turns it into audio.

High-level synthesizer Takes text and turns it into phonetic description.

SAPI Microsoft's Speech API.

Segmental Pertaining to a single speech sound.

SDK Software Development Kit.

Suprasegmental Pertaining to a domain larger than a single speech sound.

ToBI Tones and Break Indices, a framework for transcribing prosody.

TTS Text-to-Speech.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

References

Acapela Group (2005a). *Acapela Multimedia 6.0 Supported XML SAPI 5 tags*.

Acapela Group (2005b). *Acapela TTS Multimedia Version 6.00 Software Development Kit*.

AT&T (2002). *AT&T Natural Voices™Text-to-Speech Engines System Developer's Guide for the Server, Server-Lite, and Desktop Editions*. Release 1.4.

Black, A. W., P. Taylor, and R. Caley (1999). *The Festival Speech Synthesis System - System Documentation* (1.4 ed.). http://www.cstr.ed.ac.uk/projects/festival/manual/. Accessed on 4 September 2006.

Breazeal, C. (2004). Function meets style: Insights from emotion theory applied to HRI. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews 34*(2).

Bulut, M., S. S. Narayanan, and A. K. Syrdal (2002). Expressive speech synthesis using a concatenative synthesizer. In *Proceedings of ICSLP '02*.

Burkhardt, F. and W. G. Sendlmeier (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. In *Proceedings of the ISCA Workshop on Speech and Emotion*.

Cahn, J. E. (1990). Generating expression in synthesized speech. Technical report, Massachusetts Institute of Technology.

Chu, M., H. Peng, Y. Zhao, Z. Niu, and E. Chang (2003). Microsoft Mulan - A bilingual TTS system.

Edgington, M. (1997). Investigating the limitations of concatenative synthesis. In *Proceedings of Eurospeech '97*.

Eide, E., A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli (2004). A corpus-based approach to <ahem/> expressive speech synthesis. In *Proceedings of the 5th ISCA Speech Synthesis Workshop*.

Fonix (n.d.). *Fonix DECtalk Software 4.6.6 SDK Help Guide*.

Frick, R. W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin 97*(3), 412–429.

Gerkey, B., R. Vaughan, and A. Howard (2003). The Player/Stage Project: Tools for Multi-Robot and Distributed Sensor Systems. *Proceedings of the 11th International Conference on Advanced Robotics*, 317–323.

Gobl, C., E. Bennett, and A. N. Chasaide (2002). Expressive synthesis: how crucial is voice quality? In *Proceedings of the 2002 IEEE Workshop on Speech Synthesis*, pp. 91–94.

Hallahan, W. I. (1995). DECtalk software: Text-to-speech technology and implementation. *Digital Technical Journal 7*(4).

Heuft, B., T. Portele, and M. Rauth (1996). Emotions in time domain synthesis. In *Proceedings of the Fourth International Conference on Spoken Language*, Volume 3.

Hill, D. (n.d.). gnuspeech. http://www.gnu.org/software/gnuspeech/. Accessed on 4 September 2006.

Lee, C., S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan (2004). Emotion recognition based on phoneme classes. In *Proceedings of ICSLP '04*.

Loquendo (2005). *Loquendo™TTS 6.5 SDK User's Guide*.

MARY Text To Speech (n.d.). http://mary.dfki.de/. Accessed on 4 September 2006.

Microsoft (2006). Speech SDK 5.1 help file. http://www.microsoft.com/speech/SDK/51/sapi.chm. Linked to from http://www.microsoft.com/speech/techinfo/apioverview/. Accessed on 4 September 2006.

Montero, J., J. Gutiérrez-Arriola, J. Colás, E. Enríquez, and J. Pardo (1999). Analysis and modelling of emotional speech in Spanish. In *Proceedings of ICPhS 1999*, pp. 957–960.

Mozziconacci, S. and D. Hermes (1999). Role of intonation patterns in conveying emotion in speech. In *Proceedings of ICPhS 1999*, pp. 2001–2004.

Pitrelli, J., R. Bakis, E. Eide, R. Fernandez, W. Hamza, and M. Picheny (2006). The IBM expressive text-to-speech synthesis system for American English. *IEEE Transactions on Audio, Speech and Language Processing 14*(4), 1099–1108.

ProSynth Project Deliverables (n.d.). http://www.phon.ucl.ac.uk/project/prosynth/. Accessed on 4 September 2006.

Rank, E. and H. Pirker (1998). Generating emotional speech with a concatenative synthesizer. In *Proceedings of ICSLP '98*.

ScanSoft (2004). *RealSpeak Solo Software Development Kit v4.0*.

Scherer, K., D. Ladd, and K. Silverman (1984). Vocal cues to speaker affect: Testing two models. *The Journal of the Acoustical Society of America 76*, 1346–1356.

The Player Project (n.d.). http://playerstage.sourceforge.net/. Accessed on 6 September 2006.

Vroomen, J., R. Collier, and S. Mozziconacci (1993). Duration and intonation in emotional speech. In *Proceedings of the Third European Conference on Speech Communication and Technology*, pp. 577–580.

Williams, C. and K. Stevens (1972). Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America 52*, 1238.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review