On the Use of NonVerbal Speech Sounds in Human Communication

Nick Campbell[1,2]

[1] National Institute of Information and Communications Technology
[2] ATR Spoken Language Communication Research Laboratory,
Keihanna Science City, Kyoto 619-0288, Japan
nick@nict.go.jp, nick@atr.jp

Abstract. Recent work investigating the interaction of the speech signal with the meaning of the verbal content has revealed interactions not yet modelled in either speech recognition technology or in contemporary linguistic science. In this paper we describe paralinguistic speech features that co-exist alongside linguistic content and propose a model of their function and usage, and discuss methods for incorporating them into real-world applications and devices.

Keywords: interactive speech, social interaction, affect, natural data, statistical modelling, real-world applications.

1 Introduction

"Research on various aspects of paralinguistic and extralinguistic speech has gained considerable importance in recent years. On the one hand, models have been proposed for describing and modifying voice quality and prosody related to factors such as emotional states or personality. Such models often start with high-intensity states (e.g., full-blown emotions) in clean lab speech, and are difficult to generalise to everyday speech. On the other hand, systems have been built to work with moderate states in real-world data, e.g. for the recognition of speaker emotion, age, or gender. Such models often rely on statistical methods, and are not necessarily based on any theoretical models" [1].

In the fields of speech technology and multi-modal interaction, applications are already being developed from these models and data, based on published research findings and on assumed market needs. The developers of these applications might not be experts in paralinguistics or human psychology themselves, and accept the methods and assumptions of researchers in these fields as necessary and proper for the technologies. However, the data and methods required to understand basic human characteristics almost certainly do not equate to the data required to build working applications.

This paper describes some findings from an analysis of a very large corpus of spontaneous everyday conversations and shows that a considerable proportion of the speech is concerned not only with transfer of linguistic content, but also with the display of interpersonal affective information, functioning alongside, and in parallel with, the transfer of linguistic content. Whereas linguistic science and psychology may suffice to describe the content of each utterance and the various affective states of the speaker and listener,

A. Esposito et al. (Eds.): Verbal and Nonverbal Commun. Behaviours, LNAI 4775, pp. 117–128, 2007.

a new branch of communication modelling might be required to describe the social interactions per se and the pragmatic function of many of the speech sounds and speaking styles that were encountered.

Reflecting some of the more recent developments in Conversational Analysis and discourse modelling [2, 3, 4], the findings from this study have confirmed that a large amount, approximately half, of the speech sounds used in normal everyday conversational speech are nonverbal, often simply perceived as 'noise' but functioning to signal important affect-related information. However, because many of these nonverbal speech sounds are typically considered as "fillers" or "hesitations", "performance errors" (sic), or as evidence of lack of preparation of the speech utterance they are frequently edited out of recordings, disregarded in a transcription, or simply not produced at all by the professional speakers (actors, announcers, newsreaders, etc) on whom many researchers rely to produce their data for analysis.

The analysis performed on 1,500 hours of transcribed spontaneous natural spoken interactions recorded over a period of five years in the Kansai region of Japan has provided insights into new challenges for speech synthesis, new features for speech recognition, and raises questions about the need for a new grammar of spoken language which will function both independently and in conjunction with contemporary linguistic grammars. These issues will be addressed separately below.

2 A Database for Paralinguistic Research

It is imperative that any further data we collect on the same scale should be of use both to basic fundamental research issues *and* to application development. The 'Workshop Theme' of Para-Ling'07 [1] poses this question as: "How would a database need to be structured so that it can be used for both research on model-based synthesis and research on recognition?".

Those working from within the statistical approaches might respond that both synthesis and recognition benefit more from an increase in the amount of raw data than from any other type of improvement. However, this may be because both tasks have so far been constrained mainly to produce linguistic information from or for a speech signal, and because neither technology really yet tackles the psychological aspects of personal interaction and discourse management such as are signalled by changes in voice quality and prosody control.

Those working from within the social sciences might answer that a 'corpus' is of more value than a 'database', since the latter is not just a condensed and structured version of the former, but implicitly encapsulates, and is therefore limited by, current assumptions about the ontology of the subject, whereas the former also includes examples of many more features that have not yet been sufficiently understood to be included as explicit database dimensions.

Our own experience of collecting a very large amount of natural conversational speech, in the field, would not be an easy one to replicate; it was both expensive and time-consuming, and the recorded data include much information of a personal and often confidential nature so that the resulting corpus can not be readily distributed

or made publically available[1]. However, based on that experience, we do have opinions about what form a research database should take and on ways that it might be more efficiently collected.

The design constraints for collecting a representative corpus of speech should of course incorporate factors that govern size and naturalness. Given a large-enough corpus, we can assume that most *normal* aspects of interactive speech will be covered, but we can also be sure that many marginal or non-typical events will *not* be included, however large the corpus. Solving this problem requires perfecting elicitation methodologies that will provoke a natural reaction to an unnatural stimulus, and at the same time requires serious consideration about the *purpose* of the data collection, i.e., whether it is primarily to collect many examples of what a human speaker might possibly do and say (no matter how rare or unusual they may be) or whether it is to build a database of multiple examples of how they normally respond in a wide range of situations. The former is presumably the goal of the academic, the latter the goal of the engineer. The goal of the community is to establish a common ground between the two.

Labov's Observer's Paradox [5] (wherein the presence of an observer or a recording device has a measurable effect on the performance of the observed) must first be overcome in order to gather representative speech or multimodal interaction data. Furthermore, if we constrain the behaviour of our subjects in *any* way, then the results will also be unnatural, by definition. If we set any bounds at all on the data that are to be collected, then we are constraining our findings to meet our prior expectations, yet if we simply gather all and every sample that comes our way, we will be faced with some very repetitive and monotonous samples of speech. This is the Corpus-Maker's Paradox.

It is a truism that "the data define the application and the application defines the data"; a corpus that is ideal for speech synthesis may not necessarily be of any use at all for speech recognition, and vice versa. Even within the narrow confines of speech synthesis, a corpus of newsreading might be of little use for story-telling. Indeed, it may not be possible to collect all-purpose data any more than it would be reasonable to expect a single human being to be perfect at (for example) combining comedy, professional newsreading, and Shakespearean acting. Just as people specialise and develop strengths in particular areas, so the corpora we collect can only be representative of specific contexts and predetermined social situations. It is necessary first to define the purpose of the data collection.

It is suggested as part of the the workshop theme that "In application-oriented research, such as synthesis or recognition, a guiding principle could be the requirements of the 'ideal' application: for example, the recognition of finely graded shades of emotions, for all speakers in all situations; or fully natural-sounding synthesis with freely specifiable expressivity; etc. [...], and a cross-cutting perspective may lead to innovative approaches yielding concrete steps to reduce the distance towards the 'ideal'." [1]

This suggestion can be taken to imply that the defining characteristic of paralinguistics in human interaction is the (emotional) state of the speaker per se. Now, it may be that the current research needs of both psychology and linguistics can indeed be

[1] Note, however, that the ESP corpus *can* be made available, for research use only, to approved institutes and individuals subject to the signing of a non-disclosure agreement.

satisfied by facts about the speaker (or the utterance) in isolation, but the present paper argues strongly that it is instead the *common space between the speaker and the listener* that should be of most interest in terms of understanding paralinguistics for application-based research.

In 'speaker-centric' research, where different emotional states result in different lexical-choices, speaking-styles, and phrasing, the ideal corpus will be one in which the speaker experiences as many emotions of as many different varieties as possible. In 'communication-centred' research, on the other hand, while the speaker's emotional states may vary, it is the varying states of *relationships with the listener* (i.e., with the conversational partner) and the *discourse intentions of the speaker* that become more critical. It is our experience that speakers tend to monitor themselves and suppress or control display of their own emotional states during normal conversational interactions and that they focus instead on projecting an ideal state or 'character' for the current discourse purpose. They do this most obviously through prosodic modulation of feedback utterances.

3 Prosody of Paralinguistic Speech

Some novel aspects of the conversational speech encountered in the ESP Corpus will be discussed in this section. They are presented in support of the claim that at least *two* streams of information are being produced in parallel by the speaker in such interactive situations, and to argue that unless *both* streams are represented in the corpus, or simulated in laboratory data, then it will fail to be representative of typical expressive speaking styles.

The structure of spontaneous speech appears to be fragmented in much the same way as files on a computer disk can be fragmented, with individual fragments containing both inherent meaning and linking information. The discourse as a whole is made up of the combined fragments yet many of them might appear quite unintelligible in isolation. Whereas the linking information present in disk fragments is related to blocks and sectors on the disk, the linking information in a speech fragment relates it to the speaker's discourse intentions through prosody.

Continuing the computer metaphor, while the fragments on the disk are often physically separate, the files we see on the screen appear to be coherent and whole. So on listening to the speech, although we perceive a coherent stream, the phonetic transcription reveals much more fragmentation.

The entire corpus (1,500 hours of speech in all) was transcribed by hand under strict phonetic requirements: the text was to be both human-readable, *and* machine readable, accurately representing each sound that was present in the speech with some form of tag or label. Many of these sounds correspond to words in the language; about half did not (examples have been published elsewhere [6], see also [7]).

Many of the non-word sounds were laughter. In all we counted more than two-thousand types of laugh, many appearing more than a few hundred times each. Many more of these sounds were 'grunts', (equivalent to 'ummh' or 'err' in English) [8]. Others were 'wrappers', frequent phrases such as 'you, know', 'well, ...', 'let me see ...',

serving to break-up the conversation and allow the speaker to express affect through voice-quality and prosodic differences.

It is argued that the very frequent appearance of such simple nonverbal elements interspersed regularly throughout the speech allows the speaker to express not just the linguistic content, but also 'state-of-mind' through 'tone-of-voice'.

3.1 Voice Quality and Paralinguistic Speech

'Tone-of-Voice' is a term often used by the layperson but rarely by the speech professional. David Crystal uses it in "Paralinguistics" ([9], p.173) noting that "babies respond to adult tones of voice very early indeed, from around two months, and it is these which are the first effects to emerge in their own productions - from as early as seven months". Voice quality is certainly an essential part of prosody [10], though not often included in linguistically-based prosodic research or speech technology, which usually confine their interests to the 'big-three', pitch, power, and duration. In paralinguistic research, one might claim that voice-quality is *even more* important than for example segment duration or speech amplitude except in certain marked cases. Dimension-reduction experiments using Principal Component Analysis have shown for several of the speakers in the ESP data that voice-quality appears strongly in either the first or the second principal component, where the first three principal components together account for approximately half of the significant variance in the acoustic parameters of speech that is linguistically similar but functionally diverse. (For example [13, 14]).

Being a parameter that is difficult to control intentionally, voice quality serves as a strong indicator of the affective states of the speaker [11, 12], and is perhaps the most strongly recognised feature of paralinguistic speech, albeit subconsciously.

3.2 Synthesis of Paralinguistic Speech

Several approaches have been suggested to incorporate paralinguistic information in synthesised speech. While many have attempted to model the prosody of expressive speech (e.g., [15]), and even more have concerned themselves with the manipulation of voice quality parameters to distinguish between male and female voices (and very occasionally children), few have attempted to modify voice quality for paralinguistic effect.

There have also been many attempts to model 'emotion' in speech synthesis, from the work of Janet Cahn in the eighties onwards [17, 18], but almost all (see e.g., [19, 20] as notable exceptions) have concentrated on emulating the big-five (or is it six?) emotions of joy, anger, fear, disgust, sadness, and 'neutral' (sic) that have traditionally been used for research into facial expressions [21].

Our experience with the natural-conversations corpus is that such strong and marked emotions are particularly rare in everyday speech. They may present an appealing challenge to the designers of speech synthesis engines, but surely there is little call for them in real-life applications apart from story-telling and games. Much more important for business applications is the need to express interest and boredom, hesitation and politeness, warmth and distance, etc. Yet these dimensions of paralinguistic expression are seldom taken up as challenges.

3.2.1 Model-Based Approaches

Alessandro's work on voice quality modification [22] provides a strong model for the representation and modification of voice quality, where the main dimension of variation is in the range of hard/soft or breathy/pressed voice. Pressed voice being used to express enthusiasm, and creaky voice for more casual speaking styles.

Kawahara's STRAIGHT [23, 24] also provides a mechanism for voice-quality modification and has been used to replicate the expressive voice and speaking styles of Japanese Noh actors [23, 24] as well as for emotion simulation.

However, our human sensitivity to even very fine modifications of voice quality result in clear perception of any damage caused by speech signal warping and we appear to have a low tolerance to model-based speech synthesis where expression of paralinguistic information is concerned.

3.2.2 Data-Driven Approaches

Data-driven approaches, on the other hand, require very large amounts of speech data, and are strongly limited to only one voice and speaking style if high-definition, clear voice quality is a requirement.

Iida et al [25, 26] tested a multi-database approach to concatenative speech synthesis wherein a speaker was recorded for one hour each under four different emotional states and confirmed that the associated voice quality variations can be incorporated in concatenative methods.

Campbell's recent work [27, 28] also attempts to incorporate non-verbal information in the speech through use of speech segments incorporating different voice-quality characteristics. Having a five-year database of one person's speech should provide the ultimate resource for such data-driven synthesis, but in practice, we still lack a clear understanding of all the factors which control these variations and how different voice qualities will be perceived when used for synthesised utterances, so this remains as current and future work which will be reported elsewhere (see e.g. [30]).

3.3 Recognition/Classification of Paralinguistic Properties of Speech

It has proven to be particularly difficult to produce a complete and sufficient set of labels for the ESP corpus, as different labellers perceive different types of information from the same speech signal. This does not, however, imply that what they perceive is random, more that they are attuned to different dimensions of information in the signal. Taken together, the sum of all the labels describe many aspects of the speech, but individually they can be difficult to compare. For example, labeller A may determine that the speaker is 'speaking softly', Labeller B that she is 'being kind', labeller C that she is 'acting cute', and so on. Of course we can constrain the set of terms that the labellers are allowed to use to describe the data and so achieve higher 'consistency' in the labelling, but at the loss of what information?

We can instead explain the apparent confusion as follows: labeller A is being sensitive to the mechanics of the speaking style, labeller B to its pragmatic function, and labeller C more to appearance. There is no contradiction, nor any objective measure of which is more appropriate.

As well as using subjective labels of the types illustrated above, when selecting an utterance variant for concatenative speech synthesis, we also attempt to describe each speech fragment in terms of three slightly more objective dimensions. The first describes the speaker, the second her relationship with the interlocutor, and the third the intention underlying the utterance.

The speaker has at any given time various interacting states of arousal; she may have slept well, be interested in her topic, be healthy, not hung-over, etc., all of which will have an effect on her speaking style. Her relationship with the interlocutor may be close, the situation informal, relaxed, public, quiet, in a pub, etc., all having an effect on her manner of speaking. And she may be performing a greeting, in the morning, politely, etc., which three dimensions taken together collectively determine not only the manner of speaking but also the content of the utterance, its wording and complexity.

So for an ideal paralinguistic concatenative speech synthesiser, all the data would be labelled in such a way. Given five years of someone's conversational utterances preserved in a corpus, it should be feasible to synthesise most of the utterances required for the sixth year from this resource if such a general and comprehensive system of labelling could be applied to all the data. That, however, would require automatic techniques for the detection or estimation of each descriptive parameter (and if we could do that, we would have produced a very exceptional and useful computing device indeed!). This too remains as work in progress, though we now have 10% of our data manually labelled for such details.

3.4 Analysis of Paralinguistic Speech

The first stage of such automatic corpus processing requires recognition of the component speech fragments and annotation of each fragment in terms of its speaking style features. Several techniques are already available for this.

Since the speech has already been transcribed, one could suppose that further automatic labelling would be unnecessary, but that is not the case. We need finely aligned time information, at the phone level if possible, for each speech segment for prosodic analysis and speech synthesis development. Speech recognition tools that can be freely downloaded are widely available for such a task. A dictionary can be created from the transcriptions, which are also useful for training a statistical language model. With such a large amount of closed training data, recognition and alignment performance is very high.

However, as noted above, approximately half of the data is nonverbal, and the speech is also highly fragmented. Non-standard recognition is necessary in this case, for where a standard speech recogniser typically uses a set of 'garbage' models to normalise and filter out the so-called 'non-speech' noises from the speech signal, it is precisely those noises that are of most interest for use in paralinguistic feature detection.

We have therefore produced a further dictionary and language model specifically for the detection of 'grunts', laughs, and other such nonverbal speech events. Here, we treat the lexical speech (i.e., that which can be recognised well by the standard recogniser) as 'garbage', and concentrate instead on the stream of noises, detecting prosodic and voice-quality changes over time from their discrete and simple repetitions.

A dictionary of only 100 items accounts for at least half of the non-lexical speech utterances in the corpus. Our nonverbal dictionary contains several thousand items but many of them only occur very infrequently. Because the small number of common sounds (typical grunts) are so very frequent in conversational speech, these particular sounds facilitate very fine comparison of their prosodic differences. For example, when the speaker (a listener in this case) utters 'umm, umm, umm' every three seconds (which often happens in Japanese conversations), we can tell easily if she is speeding up or slowing down, if her pitch is rising across the series, or falling, if her voice is becoming relatively more or less breathy, etc., and it is from this dynamic prosodic information that our paralinguistic 'understanding' of the speech information is derived.

3.4.1 Acoustics and Physiology

The degree of tension in the voice reveals the degree of relaxation of the speaker. From the settings of the larynx, speed of the speech, range of excursion of pitch and power, etc., that is measured on the stream of nonverbal speech fragments we can form an estimate of the changing psychological and physiological states of the speaker.

We have shown in previous work [31][2] that the settings of these acoustic parameters correlate very well with differences in speaker state (e.g., the 'social' tension associated with politeness) and relationship with the interlocutor (e.g., degree of familiarity).

We have also confirmed for different speakers and for different interlocutors in a balanced conversational setup that basic voice quality settings differ consistently according to familiarity to an extent that can be reliably measured.

3.5 Assessment and Perception of Paralinguistic Speech

Returning then to the initial topic of what form an ideal database should take, we consider in this section what it is that people perceive in so-called paralinguistic speech. Or phrased differently, what aspects of the speech signal should be taken into account when evaluating a conversational utterance as suitable for inclusion in a database of speech samples?

From the above, we can conclude that it may be an oversimplification to associate paralinguistic expression simply with emotion in speech. Rather, we should consider its social function and think of it instead as an indicator of social psychological states (after Crystal, ibid, p.167, and Scherer '94 [29]). Variety in paralinguistic expression serves to indicate such interpersonal relationships as dominance, submission, leadership, and so on ... Crystal links variation in tone-of-voice with factors such as hard-sell vs soft-sell in television marketing - where the emotional state of the speaker is almost irrelevant, compared with the relationship that is being established between the speaker and the listener.

Since paralanguage serves to communicate "grammatical, attitudinal, and social information" (ibid, p.168), so a corpus for paralinguistic research and application development should be balanced not just in terms of speaker arousal, but also in terms of speaker-interlocutor relationships. If we must record such speech in a studio, then perhaps we should arrange for a series of different interlocutors to be present to motivate

[2] cf especially our "Gold-Star Slides for Science".

the speaker in different ways. Remote conversations, by telephone, are of course the easiest way to do this, without having the voice of the interlocutor interfere with the recordings of the target speaker. Having the same speaker talk in turn with friends, family members, staff, colleagues, strangers (both male and female) is the easiest way to elicit natural variation in speaking styles.

In assessing and labeling the resulting corpus, we still need to establish a framework wherein aspects of speaking style, interspeaker relationships, perceived speaker character, interpersonal stances, and so forth can be annotated and compared.

3.6 Typology of Paralinguistic Speech

Because there are not the same kinds of clear-cut distinctions between classes and types of paralinguistic information as there are between the words and phones of a language, Crystal was driven to describe paralinguistics as "the 'greasy' part of speech". One can sympathise with his frustration.

A /p/ may not gradually merge into a /b/, but interest can easily merge into boredom, and boredom into frustration. Politeness can merge gradually into familiarity, and laughter into tears. A speaker might sound cute to one listener and at the same time obnoxious to another. The categories of Paralinguistic variation cover all facets of human interaction, yet in attempting to map them we might have to include contradictory subjective descriptors as well as more objective measures based on observation of the signal.

4 Applications

How are we to reconcile this lack of a consistent framework with the concrete demands of application development? Perhaps the needs of the latter can resolve the quandries of the descriptive approach. When designing an application for human interaction, e.g., a speech interface for an advanced translation device, we can list up the situations in which it is expected to be used and precisely enumerate the capabilities required for each type of interaction it is to be built for.

But what of the ECA? Research into embodied communicative agents is now common worldwide and considerable real money is being spent by ordinary people on life in virtual worlds[3]. Here the needs are for truly expressive interactions, and in an environment unmoderated by real-world physical constraints. The expectations of the customers will be very high in such situations, and the information carried by voice quality in these very interactive environments will be (perhaps literally) explosive.

Perhaps because of the high sensitivity to voice-quality and prosody in human (or ECA) interactions it will be better to reduce realism in a way similar to graphic images in cartoons. However, it should be noted that, so far, *NO* cartoons have successfully used artificial voices alongside artificial images. All use human voices distorted to sound less like the original speaker. Perhaps this is because of the innate sensitivity of even the two-month-old baby to paralinguistic information carried by tone of voice.

[3] See e.g., "Second Life" [32] where almost six million members spent more than 1.5 million US dollars in the past 24 hours.

5 Conclusions

"Paralanguage describes the nonverbal communication that accompanies verbal communication".

This paper has presented some personal views on the use of nonverbal speech sounds in human communication, based on experiences gained from the continuing analysis of a very large corpus of spontaneous conversations. The paper has suggested that (a) conversational speech encodes two distinct streams of information, (i) linguistic, and (ii) interpersonal. Whereas the Wikipedia defines the Greek prefix 'para' as meaning 'beside', 'near', or 'alongside', this paper inclines towards the view that linguistic and paralinguistic information are intertwined rather than parallel and distinct. They coexist in much the same way that vowels and consonants coexist in speech, alternating in an irregular but well-formed way to jointly create an impression of meaning.

The paper has also suggested (b) that the so-called ill-formed, highly-fragmented nature of spontaneous speech is actually a natural evolution of the two streams, allowing content-filled linguistic fragments to be wrapped in nonverbal adjuncts that by being frequent, simple, and often-repeated, allow the listener, even one not yet personally familiar with the speaker to make fine-grained judgements about small changes in vocal settings and speech prosody.

A unifying theme of the paper has been the function of paralinguistic information, defined here as a means to clarify the speaker's attitudinal stances, to display her affective states, and to establish her relationships with the interlocutor for the purposes of the discourse, which we have modelled as three different dimensions of paralinguistic control: (a) the speaker, (b) her relationship with the interlocutor, and (c) the intention underlying the utterance. These have been tested in speech synthesis applications.

That babies of two-months can understand differences in their mother's tone-of-voice implies that this is a very basic and prelinguistic form of communication. To reduce it simply to a mere display of emotion is an oversimplification, yet we still lack words to do it justice.

Thirty-eight years ago, David Crystal closed his chapter on Paralinguistics with these words (ibid, p.174): "There is still a considerable gap, however, between our intuitive ability to recognise and interpret paralinguistic effect - our 'natural' sense of linguistic appropriateness and taboo - and our ability to state in clear terms what it is that we perceive. The spectre which still haunts papers on paralanguage, including this one, is the extraordinary difficulty of putting into words and diagrams what it is that we hear in order that the effects described be as meaningful as possible to the reader". They are clearly still true today.

Acknowledgements

This work is supported by the National Institute of Information and Communications Technology (NiCT), and includes contributions from the Japan Science & Technology Corporation (JST), and the Ministry of Public Management, Home Affairs, Posts and Telecommunications, Japan (SCOPE). The author is especially grateful to the management of Spoken Language Communication Research Labs at ATR for their continuing encouragement and support.

References

[1] Shröder, M.: Impersonal Communication: from the website of ParaLing 2007. The Workshop Theme, http://www.dfki.de/paraling07/WorkshopTheme/

[2] Allwood, J.: An activity based approach to pragmatics. Technical Report (GPTL) 75, Gothenburg Papers in Theoretical Linguistics, University of Goteborg (1995)

[3] Goffman, E.: Forms of Talk, Philadelphia, University of Philadelphia Press (1981)

[4] Shaw, M.E.: Group dynamics: the psychology of small group behaviour. McGraw Hill, New York (1981)

[5] Labov, W., Yeager, M., Steiner, R.: Quantitative study of sound change in progress, Philadelphia PA: U.S. Regional Survey (1972)

[6] Campbell, N.: How speech encodes affect and discourse information. In: Esposito, A., Bratani ć, M., Keller, E., Marinaro, M. (eds.) Fundamentals of Verbal and Nonverbal Communication and the Biometric Issue, pp. 103–114. IOS Press, Amsterdam (2007)

[7] Ward, N.: Non-Lexical Conversational Sounds in American English. Nigel Pragmatics and Cognition 14(1), 113–184 (2006)

[8] Ward, N.: Pragmatic Functions of Prosodic Features in Non-Lexical Utterances. In: Proc. Speech Prosody 2004, Nara, Japan, pp. 325–328 (2004)

[9] Crystal, D.: Prosodic Systems and Intonation in English. Cambridge University Press, Cambridge (1969)

[10] Pittam, J., Scherer, K.R.: In: Vocal expression and communication of emotion. Guilford, New York, pp. 185–197 (1993)

[11] Scherer, K.R.: Vocal affect expression: A review and a model for future research. Psychological Bulletin 99, pp. 143–165 (1986)

[12] van den Brock, E.L.: Empathic Agent Technology (EAT). In: Johnson, L., Richards, D., Sklar, E., Wilensky, U. (eds.) Proceedings of the AAMAS-05 Agent-Based Systems for Human Learning (ABSHL) workshop, pp. 59–67. Utrecht, The Netherlands (2005)

[13] Campbell, N., Mokhtari, P.: Voice Quality; the 4th prosodic parameter. In: Proc 15th ICPhS, Barcelona, Spain (2003)

[14] Campbell, N., Nakagawa, A.: 'Yes, yes, yes', a word with many meanings; the acoustics associated with intention variation. In: Proc ACII 2007 (Affective Computing and Intelligent Interaction) Lisbon, Portugal (2007)

[15] Hamza, W., Bakis, R., Eide, E.M., Picheny, M.A., Pitrelli, J.F.: The IBM Expressive Speech Synthesis System. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP), Jeju, South Korea (October 2004)

[16] Pitrelli, J.F., Bakis, W., Eide, R., Fernandes, E.M., Hamza, R., Picheny, M.A.: The IBM Expressive Text-to-Speech Synthesis System for American English. IEEE Transactions on Audio, Speech, and Language Processing 14(4), 1099–1108 (2006)

[17] Cahn, J.E.: Generating expression in synthesized speech. Master's thesis, Massachusetts Institute of Technology (1989), http://alumni.media.mit.edu/~cahn/emot-speech.html

[18] Cahn, J.E.: The Generation of Affect in Synthesized Speech. Journal of the American Voice I/O Society 8, 1–19 (1990)

[19] Trouvain, J., Schroeder, M.: How (not) to add laughter to synthetic speech. In: Proc. Workshop on Affective Dialogue Systems, pp. 229–232. Kloster Irsee, Germany (2004)

[20] Schroeder, M.: Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. In: Proc. Workshop on Affective Dialogue Systems, Kloster Irsee, Germany. LNCS, pp. 209–220. Springer, Heidelberg (2004)

[21] Ekman, P.: Universals and cuntural differences in facial expression of emotion. In: Cole, J.K. (ed.) Nebraska Symposium on Motivation, pp. 207–282. University of Nebrasaka Press, Lincoln (1972)

[22] d'Alessandro, C., Doval, B.: Voice quality modification for emotional speech synthesis. In: Proc. Eurospeech 2003, Geneva, Switzerland, pp. 1653–1656 (2003)
[23] Kawahara, H., de Cheveigné, A., Banno, H., Takahashi, T., Irino, T.: Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT. In: Proc. Interspeech 2005, Lisboa, pp. 537–540 (2005)
[24] Modification of Japanese Noh voices for speech synthesis: http://www.acoustics.org/press/152nd/kawahara.html
[25] Iida, A., Campbell, N., Yasumura, M.: Design and Evaluation of Synthesised Speech with Emotion. Journal of Information Processing Society of Japan 40 (1998)
[26] Iida, A., Campbell, N., Iga, S., Higuchi, Y., Yasumura, Y.: A speech synthesis system with emotion for assisting communication. In: Proceedings of the ISCA Workshop on Speech and Emotion, Belfast, pp. 167–172 (2000)
[27] Campbell, N.: Specifying Affect and Emotion for Expressive Speech Synthesis. In: Gelbukh, A. (ed.) CICLing 2004. LNCS, vol. 2945, Springer, Heidelberg (2004)
[28] Campbell, N.: Conversational Speech Synthesis and the Need for Some Laughter. IEEE Transactionson Audio, Speech, and Language Processing 14(4), 1171–1179 (2006)
[29] Scherer, K.R.: Interpersonal expectations, social influence, and emotion transfer. In: Blanck, P.D. (ed.) Interpersonal expectations: Theory, research, and application, pp. 316–336. Cambridge University Press, Cambridge and New York (1994)
[30] Campbell, N.: Expressive / Affective Speech tone-of Synthesis. In: Benesty, J., Sondhi, M.M., Huang, Y. (eds.) Springer Handbook on Speech Processing and Speech Communication, Springer, Heidelberg (2007) (in press)
[31] Campbell, N.: Getting to the heart of the matter; speech as expression of affect rather than just text or language. In: Language Resources & Evaluation, vol. 39(1), pp. 109–118. Springer, Heidelberg (2005)
[32] Second Life: a 3-D virtual world entirely built and owned by its residents. Since opening to the public in, it has grown explosively and at the time of writing is inhabited by a total of 5,788,106 people from around the globe (2003), http://secondlife.com/