Specifying Affect and Emotion for Expressive Speech Synthesis

Nick Campbell

ATR Human Information Science Laboratories, Kyoto, Japan.
nick@atr.co.jp

Abstract. Speech synthesis is not necessarily synonymous with text-to-speech. This paper describes a prototype talking machine that produces synthesised speech from a combination of speaker, language, speaking-style, and content information, using icon-based input. The paper addresses the problems of specifying the text-content and output realisation of a conversational utterance from a combination of conceptual icons, in conjunction with language and speaker information. It concludes that in order to specify the speech content (i.e., both text details and speaking-style) adequately, selection options for speaker-commitment and speaker-listener relations will be required. The paper closes with a description of a constraint-based method for selection of affect-marked speech samples for concatenative speech synthesis.

1 Introduction

For unrestricted text-to-speech conversion, the problems of text anomaly resolution and given/new or focus determination can be profound. They can require a level of world-knowledge and discourse modelling that is still beyond the capability of most text-to-speech synthesis systems. One implication of this is that the prosody component of the speech synthesiser can only be provided with a default specification of the intentions of the speaker or of the underlying discourse-related meanings and intentions of the utterance, resulting in a flat rendering of the text into speech. This is not a problem for the majority of synthesis applications, such as news-reading or information announcement services, but if the synthesiser is to be used in place of a human voice for interactive spoken dialogue, or conversation, then the speech will be perceived as lacking in illocutionary force, or worse, it will give the listener a false impression of the intentions of the utterance and of the speaker-listener relationships, leading to potentially severe misunderstandings.

When a synthesiser is to be used in place of a human voice in conversational situations, such as in a communication aid for the vocally impaired, in speech translation systems, or in call-centre operations, then there is a clear need for the vocal expression of more than just the semantic and syntactic linguistic content of the utterance. Paralinguistic information related to dialogue turns, and speaker interest is signalled along with the syntactic structure of the speech by means of prosody and voice quality [1].

Since the information signalled in human speech includes linguistic, paralinguistic, and extra-linguistic layers, the listener presumably parses all three sources to gain access to the intended meaning of each utterance. Just as stereoscopic vision yields more than the simple sum of input from the two eyes alone, so paralinguistic speech understanding gives us more than just the sum of the text and its prosody alone [2]. For example, the lexical item 'yes' doesn't always function to mean *yes* in conversation; when spoken slowly and with a rise-fall-rise intonation, it can instead be interpreted as meaning 'no', or as signalling hesitation (i.e., that the premise is understood but not ncessarily agreed to), thus paralinguistically qualifying the interpretation of the lexical content and signalling both speaker-affect and discourse-related functions to the listener. Someone speaking with 'an authoritative tone of voice' is more likely to be listened to! Similarly, if it is clear from the speaking style that a speaker is intoxicated (for example) then the listener may be likely to interpret the content of that speech with more caution. If it is apparent that this style of speaking is under conscious control, then the words can take on yet another meaning. Such speaking-style information is not yet freely controlled by the current generation of speech synthesisers.

Paralinguistic information, signalled by tone-of-voice, prosody, and speaking style selection, becomes more important as the conversation becomes more personal. Newsreaders and announcers can distance themselves from the content of their utterances by use of an impersonal 'reporting' style of speaking, but customer-care personnel may want to do the opposite, in order to calm a client who is complaining, or to reassure one who is uncertain. When speaking with friends, for example, we use a different speaking style and tone-of-voice than when addressing a stranger or a wider audience. Voice quality is controlled (though probably not consciously) for politeness, for interlocutor, and for different types of speech act [1]. Speech synthesis must become capable of expressing such differences if it is to be of use in personal or conversational applications.

2 Expressive Speech

As part of the JST (Japan Science & Technology Agency) CREST (Core Research for Evolutional Science and Technology) ESP (Expressive Speech Processing) Project [3, 4], we are collecting 1000 hours of interactive daily-conversational speech, and are building an interface for a CHATR-type synthesiser [5, 6] to allow synthesis of speech from the resulting corpus that will be capable of full expressive variation for paralinguistic effect.

Volunteers wear head-mounted close-talking studio-quality microphones and record their daily spoken interactions to Minidisc devices in blocks of 160 minutes each [7, 8]. We now have more than two-years worth of such daily-conversational speech data from a small number of subjects. These samples are transcribed manually and segmentally aligned automatically from the transcriptions. A large part of the research effort is concerned with the choice of appropriate features for describing the salient points of this interactive speech, and with the development

of algorithms and tools for the automatic detection and labelling of equivalent features in the acoustic signal [9, 10].

Part of this project includes the development of a communication aid [11, 12] and, in particular, an interface for the speedy input of target utterances (the subject of the first part of this paper). We are not concerned with text-to-speech processing in this project, and require instead a fully annotated input that is rich enough to specify not just the lexical content of the desired utterance, but also the speaking style (including the paralinguistic and extralinguistic features) so that the synthesised speech will match the discourse context and enable the 'speaker' to convey all aspects of the intended meaning.

We have been testing our prototypes with disabled users, including muscular-dystrophy or ALS patients, who need a speech synthesiser for essential daily communication with friends, family, and care providers [13], but we also envisage business and other uses of such a system in situations where overt speech may be difficult. For example, a busy executive may want to telephone home to inform her partner that she will be returning later than usual because of an unexpectedly lengthy business meeting. She might prefer to use a synthesiser to speak on her behalf, in order not to disturb the meeting. She may also want to convey information regarding the progress of the business deal at the same time. In such a case, the words 'I'll be late tonight' could be spoken (synthesised) with a happy voice to indicate that positive progress is being made. However, if the same message were intended as warning or as an apology, then a happy voice would be quite inappropriate. As noted above, human listeners read as much from the tone of voice and speaking style in such cases as they do from the linguistic message. [1]

The JST/CREST ESP project aims at producing synthesised speech that is able to express paralinguistic as well as linguistic information, and from our analysis of the data collected so far (about 250 hours of transcribed speech, with the same amount yet to be transcribed for this collection paradigm) we observe that as the interactions become more personal, so the paralinguistic component takes on a greater role in the speech. Utterances become shorter, more common knowledge is assumed, and prosody and voice-quality carry a larger proportion of the information in the message; i.e., the speech becomes more expressive. This is in contrast to most of the corpora used for speech synthesis, where a trained or professional speaker ususally reads from prepared texts in order to produce a balanced corpus with the least amount of effort [14]. We were surprised to find that more than half of the speech, in terms of the number of utterances transcribed, consisted of non-lexical items or 'grunts', for which no dictionary entry exists, and which can only be interpreted in terms of discourse control and expression of speaker affect by their prosody and phonation characteristics. These may turn out to be the most difficult speech items to synthesise, because they are textually ambiguous and require paralinguistic descriptors for their specification.

[1] Animals, on the other hand, and soon robots too, may actually read more from the tone of voice than from the content of the speech.

Fig. 1. A sample screen-dump of the GUI interface for use from a web page or personal assistant (left), and a Java-based (i-mode) interface downloaded to a cell-phone (right)

3 Icons and Utterances

In the case of the business user described above, the use of a keyboard for inputting the text would be highly intrusive into the social situation of a business meeting. Annotating that text for paralinguistic and speaking-style information would also be a tedious and time-consuming process. For such situations, we have designed a front-end interface to the synthesiser, for use with a personal assistant or cell phone, so that the speaking style and message can be selected quickly from a menu by toggling buttons to choose between iconic specifiers for the selection parameters. Figure 1 shows a sample screen-dump of the GUI interface, programmed in Flash with a socket-based perl interface to CHATR, for use from a web page or personal assistant (left), and the equivalent Java-based i-mode interface, downloaded to a cellular phone (right).

3.1 Speech Content Specification

Because this device is intended not for the synthesis of unrestricted text-to-speech, but primarily for the generation of interactive daily-conversational utterances, we take advantage of the repetitive and simple nature of this speaking style. Most of the utterances in daily conversation are heavily stylised and repetitive, and the conversations are made up of novel combinations of these basic forms. However, while these building-blocks of conversation may be textually simple and limited in number (many of them are backchannel utterances, laughs, grunts, and fillers), their prosodic realisation can be complex and varied.

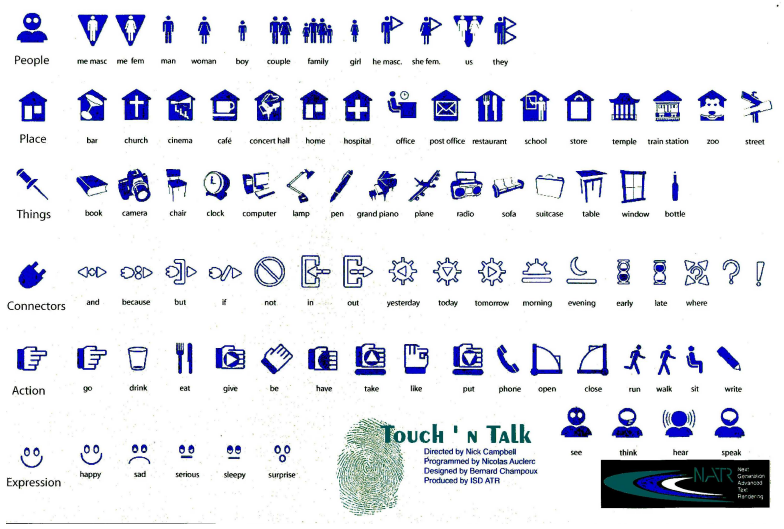

Fig. 2. A chart of the icons currently being tested for use in text selection - objects, operators, and connectives are included, but the list is not intended to be exhaustive at this early stage

We therefore designed this prototype synthesiser to enable the rapid generation of text strings with facilities for the user to easily specify the affectual flavourings and discoursal function of each utterance.

The buttons on the right of the display in figure 1 are used for selecting speaker, speaking-style, and language respectively. They toggle to show the owner of the voice (male, female, asiatic, caucasian, young,old, etc.), the emotion desired for the utterance (currently happy, sad, angry, and 'normal'), and a flag indicating the language (currently only Japanese and English [19, 20] though dialectal variants of each are available). The equivalent functions on the cell-phone are bound to the numeral keys 1, 2, and 3 on the dial pad. The icons mapped to the text buttons (left of the main display in figure 1) are illustrated in Figure 2 in the form of a table. Utterance-content icons are represented by their equivalent text on the cell-phone screen. These are bound to the numeral keys 7, 8, and 9 on the cell-phone. 0 is mapped to the 'enter' function to activate the synthesiser. The synthesised speech can be sent directly to the user's device, or redirected to a distant phone. The text icons are grouped into five functional classes: 'people', 'places', 'things', 'actions', and 'connectors'.

By selecting a combination of these icons, the text to be synthesised can be specified. A simplified version of the text, indicating key words alone, appears for confirmation in the central display window (the sand-box) and can be edited if required. A separate window can be popped up for the entry of additional items from a customisable user-specified word list, e.g., for proper names or personalised slot-fillers. This minimal iconic specification of the utterance (subject,

verb, object(s), conective(s), and modifier(s)) allows for automatic adjustments to the final wording of the text according to language, speaker, and speaking-style settings (and according to the known limitations of the synthesiser). The choice of speaker can be programmed to change voice, formality, or personality of the selected speaker, with subsequent effects on the wording, prosody, and pronunciation of the utterance.

3.2 Speaking Style Specification

The texts of all the utterances to be synthesised are produced from elementary components stored in the device (or on the server in the case of cell-phone access) as in domain-specific synthesis. They are finite in number and can be associated with parameter tables specifying e.g., breathiness of the voice, pitch inflections, durational lengthening etc., according to the combination or selection of other parameters by the user.

In the first prototype implementation of this interface, when the user selects an emotion icon, the settings for the speaker-database are changed, and the speech is synthesised using separate source databases, each characterising a different emotion. Work is in progress both to merge these individual 'emotionally-marked' databases for each speaker, to enable selection using higher-level descriptors of the speech-style characteristics, and to replace the hard-wired database-switching with an improved expressive unit-selection procedure using the large conversational-speech corpora, as detailed below.

The final text generation is hard-coded using a series of conditional and branching operations. All combinations of frequently-used components are exhaustively listed in the source code, and the appropriate prosodic and speaking-style annotations are then added manually. This step is both inelegant and labour-intensive, and we are considering methods of automating the creation of the dictionary component from an analysis of the transcriptions in the ESP corpus to take advantage of the repetitive nature of conversational speech. However, because the text, the translation, the prosody, the voice characteristics, and the speaking style can be all pre-programmed, and do not need to be computed by the synthesiser at run-time, a higher quality of synthesised speech can be guaranteed. The problems of the text-processing and prosody-prediction components have been eliminated from the synthesis process and the brunt of the responsibility now rests on the unit-selection procedures. Furthermore, the inflections of the text can be adjusted to produce an utterance that is fitting to the selection of speaker and style.

3.3 Emotion and Style

Experience with testing the above interface has revealed several aspects of the design that need further consideration. In addition to the database merging and dictionary automation mentioned above, we must also consider changes to the 'emotion' (speaking-style) selector. The interface was prepared before we had started analysing any speech from the conversational corpus, and was designed

primarily to facilitate the expression of emotion in synthesised speech. However, analysis of the conversational-speech corpus in terms of emotion, using the broad-class labels 'happy', 'sad', 'angry', and 'normal' has proved extremely difficult, for several reasons.

Firstly, the definition of 'normal' appears to be highly context-dependent, as the speaking style varies according to both familiarity with the interlocutor, and type of conversation. By far the majority of the speech falls under this category, and there are remarkably few angry or sad tokens in the corpus (which now contains more than two-years of speech). Normal seems to be 'moderately happy', but rather than expressing pure emotion (which is perhaps just an extra-linguistic aspect of the speech, irrelevant to the discourse), 'speaker involvment' and 'discourse intention' appear to be the main dimensions of paralinguistic variation. Many of the extracts that we examined (often just one side of a phone conversation with a friend or relative) were textually very repetitive, but prosodically extremely rich, and varied considerably in their expressivity and functional meaning. Much of the 'language' consisted of grunts and fillers, often monosyllabic, or repeating the same syllable many times. There is no facility for such back-channelling in the current interface, nor any way of specifying the 'flavour of the grunt' if there were.

Secondly, the 'emotion' labels have proved to be over-simplistic. It is not at all easy to classify a given utterance into one of the above basic classes without first making clear whether we are referring to the speaker's subjective emotional states (both short-term, and long-term) or to the emotional colouring of the utterance itself (and whether intended or not). A dimension of 'control' is needed in addition to the switch for emotion, so that we can distinguish between revealed and intended variants. For example, a schoolteacher might not in fact be angry when speaking (as part of the job) in an angry manner to control unruly students in the class. Conversely, the person might be feeling extremely angry (for unrelated personal reasons), but manages for social reasons not to reveal it in the speech. Both of these variants are marked with respect to speaking style.

For the labelling of paralinguistic characteristics in the speech database, each utterance must be evaluated separately in terms of such features as the relationships between speaker and hearer (age, sex, familiarity, rank, politeness, etc.), the degree of commitment to the content of the utterance (citing, recalling, revealing, acting, informing, insisting, etc.), the long-term and short-term emotional and attitudinal states of the speaker, the pragmatic force of the speech act, the voice-quality of the utterance (breathy, relaxed, pressed, forced), and so on. The list is not complete. The simplistic notion of a single switch for 'emotion' in an expressive speech synthesiser would appear to need considerable rethinking. The reduction of such complex features to a simple descriptor continues as the core of our work.

3.4 Future Work

In place of an over-simplistic 'emotion button' on the synthesiser interface, we are now considering a combination of three 'buttons', or feature dimensions, for

determining the paralinguistic information that governs the speaking style of an utterance; one for 'self', one for 'other', and one for 'act'. The 'self' button might be toggled between four states, to select between two levels (high and low) of interest and mood. The 'other' button similarly with two levels (high and low) of 'friend' and 'friendly' relationship to the listener (this will be the first time that a difference in the listener will be considered as a factor in speech synthesis output control). The 'act' (or illocutionary-force) button will be used to specify directionality of the utterance (whether offering or eliciting) and its intention (conveying primarily either affect or information).

It is clear from our observations of the ESP conversational speech corpus that, rather than a single 'emotion' factor, at least these three dimensions of speaking-style control are required. The speaker's degree of involvement in the utterance is of prime concern — the amount of interest in the topic of the conversation, the quality of personal experience underlying the expression of the current utterance, and the degree of belief in the premise of the utterance, as well as factors such as the speaker's current mood and state of health. For simplicity we have reduced these parameters to two levels, high or low, of interest and mood.

Relationship with the listener appears to be the second most important determinant of speaking style. Whether talking to a friend or a stranger, a businesss acquaintance or a family member, and whether the context of the discourse allows a more or less friendly or intimate speaking style. Simply knowing the relationship with the listener is not enough; it is also important to know the circunmstances in which the discourse is taking place. For example, the speaking style adopted when answering a question by a family member may depend on whether that question is asked at the dining table or in a conference hall.

Thirdly, the intentions underlying the utterance, the pragmatics of the speech event, play an important part in determining how it is to be realised. There are many sections of a conversation where the content (and intent) could be fully specified by a simple transcription of the words alone, and where prosody plays only a small part. We denote these as 'information'. And there are almost as many sections where the words themselves play a smaller part than the way in which they are said — we denote these as 'affect'. For simplicity, we choose to consider for our next implementation only the directionality and type of each utterance — whether the speech event functions primarily to express (or to request the expression of) affect or information. It must remain as future work to determine how the combination of each is realised in a single utterance.

4 Unit Selection for Expressive Speech Synthesis

This section describes a method by which the synthesis of spontaneous-sounding conversational speech can be generated using a small speech synthesiser as a driver for concatenative unit selection from a large spontaneous speech database. The driver is used to generate acoustic targets, which are then used for pre-selection of acoustic waveforms for the synthesis units, with the final candidates

filtered according to further acoustic constraints to ensure the selection of units having appropriate voice and speaking style characteristics.

Mokhtari & Campbell [15] have described a method whereby acoustic syllables can be automatically demarcated in running speech and used as targets for the selection of units of natural speech from a large database for concatenative synthesis. Here, we show how those targets can be used as the basis of spontaneous or expressive speech synthesis, further selecting from among the candidates thus obtained by use of voice-quality, speaking-rate, and pitch-range for a finer control of speaking styles.

There are now several very large corpora of spontaneous speech that could be used as resources for producing spontaneous-sounding speech synthesis (e.g., [16–18]) but the task of segmentally labelling them is considerable, and many of the problems related to modelling the acoustic characteristics of spontaneous speech have yet to be resolved. Until recently, detailed phonemic labelling has been a prerequisite for the use of a database in concatenative speech synthesis. However, previous work on the synthesis of multilingual speech [19, 20] resulted in a procedure for using the voice of a speaker of the target language to generate a sequence of acoustic vectors that can then be used for the selection of units from the native-language database of a non-native speaker of the target language. This work was carried out in the framework of multilingual synthesis for speech translation and resulted in more natural pronunciation of the target 'foreign language' speech, but can be applied as well to the synthesis of spontaneous or conversational speech. i.e., we can use a subset of well-labelled speech, in the voice of a given speaker, as targets for the selection of units from a larger database of spontaneous speech from the same speaker.

4.1 Corpus-Based Expressive Speech Synthesis

As noted above, by far the majority of databases that have been used for speech synthesis research have been purpose-designed and carefully read, usually by professional announcers, under studio conditions. They are not representative of 'speech in action', nor do they include the variety of natural speaking styles and situations that are encountered in everyday spoken conversations. Such studio recordings of speech can illustrate many of the formal linguistic aspects of spoken language, but few of the functional social or interactive aspects of spoken communication. Many attempts at producing corpora of spontaneous speech for synthesis research have failed, due to the acoustic and psychological difficulties of capturing natural samples of ordinary speech in everyday interactive situations. The "Observer's Paradox" (after Labov [21]) is well known to researchers in the social sciences, who have observed that when people are confronted with a microphone, their speech undergoes subtle changes and may no longer be representative of that used in their normal daily-life interactive situations. Several ways have recently been proposed to overcome this problem.

For the Corpus of Spontaneous Japanese [16] the researchers have chosen to collect a subset of spontaneous speech limited to those situations where a microphone is a common and predictable accessory, such as public lectures and

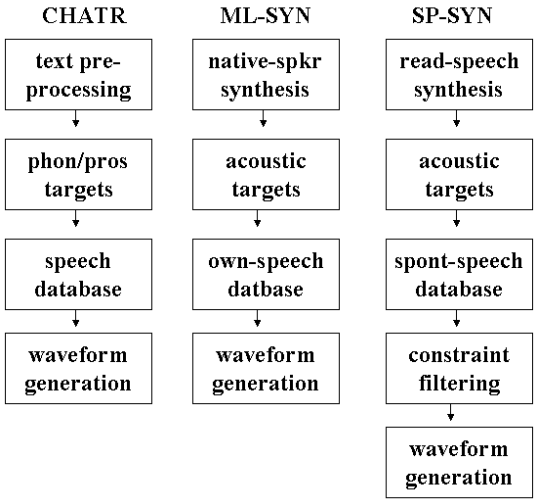

Fig. 3. Comparing the three methods of synthesis. The left column is common to all three systems, and is standard for CHATR. The middle column shows the flow for multi-language synthesis (ML), taking input (top box) from the output of the left column (bottom box) to use a native-speaker's voice as target. The right column shows the proposed system, similar to ML, but with the added filtering necessary to reduce the diversity of candidates found in the spontaneous speech database.

broadcasting environments. The resulting speech is spontaneous but formal, i.e., the speakers show a public rather than a personal face. The JST/CREST ESP corpus is another example, where the speakers and their interlocutors become accustomed to the presence of a microphone by having it worn on the body for extended periods of time. Both the above projects have produced speech data that is illustrative of the type that will be used in future corpus-based speech synthesis, but these corpora raise many problems that were not encountered in traditional concatenative speech synthesis.

The biggest problem, perhaps, lies in the amount of manual work that has been required for the processing of a speech database in order to produce a set of units suitable for concatenation. As corpora become larger, this must be automated or reduced. Consequently, we have proposed a method for using acoustic vectors as targets for the unit-selection [9, 10, 15]. The acoustic parameters obtained by concatenative synthesis from a relatively restricted (typically read-speech) database of the source speaker, are used as targets for selecting from among the loosely-matching candidates in the much larger non-read (spontaneous) speech database of the same speaker. This method relies upon the fact that the spectral representation of speech varies less than the possible prosodic representations of an equivalent utterance. We can select all acoustic waveforms having the same phonemic segmental content as the target utterance and then

further select from amongst them acording to the desired prosodic and voice-quality characteristics.

This paralinguistic speaking-style-based constraint on the unit-selection procedure requires a third stage of unit-selection to filter the candidates according to stylistic attributes that are more varied in the spontaneous speech corpus than in the read speech corpus (see Figure 3). The acoustic targets which correlate with the phonetically-motivated vocal-tract configurations (such as formants) are weighted more heavily as initial selection criteria than those which correlate with the prosodic aspects of the speech (such as spectral tilt). The prosodic constraints (e.g., 'high breathiness', 'low fundamental frequency', 'slow speech') are then used as filters in an intermediate stage to reduce the number of candidates to only those having the appropriate speaking-style characteristics for the desired 'spontaneous' speech utterance. The final selection stage is further constrained by 'join-cost' concatenation criteria to ensure a smooth and continuous sequence of units for signal generation.

5 Conclusion

For conversational speech synthesis, the text of an utterance alone is inadequate as a specification of the style in which it is to be rendered, which depends largely upon interacting factors such as speaker-state, purpose, involvement, and relations with the listener.

This paper has presented techniques for the synthesis of conversational speech, expressing paralinguistic information by means of pre-stored annotations on texts. Variants are selected by a combination of choices that define the basic components of each utterance, such as voice, language, and speaking style. The prototype is still rudimentary, but experience with this interface is allowing us to consider better ways of simply specifying the attributes of speech to be synthesised.

The paper has also presented a novel method for the synthesis of spontaneous-sounding speech, using a smaller read-speech database from the same speaker as a bootstrap for producing intermediate acoustic targets. The best-matching candidates are selected from the spontaneous-speech database and filtered by prosodic constraints so that only those having the appropriate speaking-style characteristics will be made available to the waveform-concatenation module.

Acknowledgements

This chapter is an extended and updated version of work first presented at the Acoustical Society of Japan and the 2001 IEEE Speech Synthesis workshop, with contributions from P. Mokhtari and researchers and staff of the ESP project, N. Auclerc, B. Champoux, and students of the Department of Applied Linguistics at NAIST. This work is supported partly by a grant from the Japan Science & Technology Agency under CREST Project #131, and partly by aid from the Telecommunications Advancement Organisation of Japan.

References

1. Campbell, N & Mokhtari, P., "Voice Quality; the 4th prosodic parameter", in Proc 15th ICPhS, Barcelona, Spain, 2003.
2. Antoine Auchlin, Linguistics, Geneva, personal communication, 2003.
3. JST/CREST Expressive Speech Processing project, introductory web pages at: http://feast.his.atr.co.jp/
4. Campbell, W.N., "Databases of Emotional Speech", in Proc ISCA (International Speech Communication Association) ITRW on Speech and Emotion, pp. 34-38, 2000.
5. Campbell, W. N. and Black, A. W. "CHATR a multi-lingual speech re-sequencing synthesis system". Technical Report of IEICE SP96-7, 45-52, 1996.
6. Campbell, W. N. "Processing a Speech Corpus for CHATR Synthesis". Proceedings of The International Conference on Speech Processing pp.183-186, 1997.
7. Campbell, W. N., "The Recording of Emotional speech; JST/CREST database research", in Proc LREC 2002.
8. Campbell, N & Mokhtari, P., "DAT vs. Minidisc — Is MD recording quality good enough for prosodic analysis?", Proc ASJ Spring Meeting 2002, 1-P-27.
9. Campbell, W. N., Marumoto, T., "Automatic labelling of voice-quality in speech databases for synthesis", in Proceedings of 6th ICSLP 2000, pp. 468-471, 2000.
10. Mokhtari, P, & Campbell, W. N., "Automatic detection of acoustic centres of reliability for tagging paralinguistic information in expressive speech.", in Proc LREC 2002.
11. Iida, A., Iga, S., Higuchi, F., Campbell, N., Yasumura, M., "A speech synthesis system with emotion for assisting communication", ISCA (International Speech Communication and Assosiation) ITRW on Speech and Emotion, pp.167-172, 2000.
12. Iida, A., Campbell, N. and Yasumura, M. "Design and Evaluation of Synthesised Speech with Emotion". Journal of Information Processing Society of Japan Vol. 40, 1998.
13. Iida, A., Sakurada, Y., Campbell, N., Yasumura, M., "Communication aid for non-vocal people using corpus-based concatenative speech synthesis", Eurospeech 2001.
14. Campbell, W. N., "Recording Techniques for capturing natural everyday speech", in Proc Language Resources and Evaluation Conference (LREC-2002), Las Palmas, Spain, 2002.
15. Mokhtari, P. and Campbell, N. "Automatic measurement of pressed/breathy phonation at acoustic centres of reliability in continuous speech", in Special Issue on Speech Information Processing of the IEICE Transactions on Information and Systems, The Institute of Electronics, Information and Communication Engineers, Japan, Vol. E-86-D, No. 3 (March), pp.574-582, 2003
16. Maekawa, K., Koiso, H., Furui, S., & Isahara, H., "Spontaneous Speech Corpus of Japanese", pp 947-952, Proc LREC 2000, Athens, Greece, 2000.
17. Switchboard telephone-speech database: www.ldc.upenn.edu.
18. CALLFRIEND: a telephone-speech database, LDC Catalog, 2001.
19. Campbell, W. N., "Foreign-Language Speech Synthesis", Proceedings ESCA/COCOSDA 3rd Speech Synthesis Workshop, Jenolan Caves, Australia 1998/11/26.
20. Campbell, W. N., "Multi-Lingual Concatenative Speech Synthesis", pp.2835-2838 in Proc ICSLP'98 (5th International Conference on Spoken Language Processing), Sydney Australia 1998.
21. Labov, W., Yeager, M., & Steiner, R., "Quantitative study of sound change in progress", Philadelphia PA: U.S. Regional Survey, 1972.