# Databases of Expressive Speech

Nick Campbell
ATR Human Information Science Laboratories
JST/CREST Expressive Speech Processing Project
Kyoto 619-0288, Japan
E-mail: nick@atr.co.jp

## Abstract

*This paper discusses the construction of speech databases for research into speech information processing and describes a problem illustrated by the case of emotional speech synthesis. It introduces a project for the processing of expressive speech, and describes the data collection techniques and the subsequent analysis of supra-linguistic, and emotional features signalled in the speech. It presents annotation guidelines for distinguishing speaking-style differences, and argues that the focus of analysis for expressive speech processing applications should be on the speaker relationships (defined herein), rather than on emotions.*

## Keywords

*speech synthesis, unit-selection, prosody, voice-quality, attitude, emotion, speaker-relationships.*

## 1. Introduction

Because current speech databases are custom-designed to illustrate specific aspects of speech, they may fail to adequately represent the speech of the common person as used in normal daily conversations. There has been considerable attention paid recently to the construction and evaluation of language resources [1,2], and without well-constructed text and speech databases it would be difficult to achieve technological advances in the areas of speech and language processing. There are now many excellent databases available to researchers in Japan, and the contents of these databases will govern the directions of future speech technology research. The challenges to speech recognition and speech synthesis in particular are defined by the types of speech data made available.

Spoken language information includes not just linguistic but also paralinguistic and extralinguistic components, and the challenge facing current speech technology is to incorporate these higher levels of information by parameterising them separately and modelling their interactions with the linguistic component of the message. Current speech databases are limited in their ability to illustrate supra-linguistic speech variation because of the constraints on corpus design which will be discussed below. This paper suggests a method of speech data collection that will overcome the design problem and will result in more expressive speech samples.

## 2. Speech Databses

The speech databases in current use have been designed to represent the challenges that the researchers of the time considered as most important for their incremental progress. This 'top-down', researcher-driven, approach to technological advancement through database design may be limited by the imagination of the designers, or by the needs of the industry, and there is no formal process to guarantee that it will evolve in an optimal direction. A bottom-up alternative to database design would require a different type of approach, based not on the needs of the researchers, but on the habits of the user community, as defined by their everyday language use.

Historically, speech databases have grown in size, and have evolved from being controlled examples of the acoustic-phonetic characteristics of the basic speeech sounds recorded in e.g., triphone contexts, through examples of isolated and spelled words, monosyllables, city-names, numerals, and control-words, to sets of phonemically-balanced sentences, newspaper texts, broadcast news, and sets of prompts designed to illustrate the various pronunciations of what were considered (predominently by the telecommunications industry) to be characteristic samples representative of human speech.

More recently, the speech content has become less controlled, including telephone conversations between friends, children's speech, and interactive dialogue speech (see for example the catalogues in [1,2]). The latest corpus initiative in Japan [3] is focussing on prepared spontaneous speech produced in a monologue context, especially including conference and workshop presentations. However, there is as yet no database representing the speech of the 'man or woman in the street' in normal everyday conversation. We have many examples of humans talking, but few illustrating speech in a natural environment.

When recording speech corpora as a source of units for waveform-concatenatiion speech synthesis [4,5], we prepared balanced texts that ensured even coverage of all phone combinations in most prosodic environments, but the resulting sentences, being generated by a greedy algorithm, were lexically dense and phonetically complicated for the reader to produce. The consequent stress in the reader's voice remained throughout the speech synthesis process, and the results were less than satisfactory to listen to.

Developments in synthesis corpus segmentation techniques allowed use of longer texts, such as novels or short stories, which had simpler and sequentially-related sentences but which, when read for an equivalent amount of time, provided similar prosodic and phonemic balance to the previously-used sentence lists [6]. The stories, however, being more interesting to read, had the advantage of producing a much more relaxed and fluent speaking style.

The resulting corpora were both natural-sounding and phonetically/prosodically balanced, but they were limited in that they exhibited the characteristics of only one fixed speaking style. If the source text was sad, for example, then the whole corpus would be read in a sad voice, and any synthesis produced using that speech would also sound sad because of the characteristics of voice quality and prosody in the source database. It soon became obvious that synthesis of e.g., a weather forecast made using a sad voice introduced a subjective, paralinguistic, level of interpretation of the text that could be quite different from what was intended. This prompted us to study emotional speech; not so that the synthesiser should sound emotional, but so that the synthesised speech would sound appropriate to the content and context of the target utterance

There has been growing interest recently in the topic of emotional speech and its applications in speech technology [7,8], but in this paper we question whether 'emotion' is the most appropriate aspect for spoken communication, since a speaker's *state-of-mind*

should be of less interest than his or her expressed *intentions* and *relationships* with respect to the discourse content and context. For spoken language processing, we need to know more about how the speaker relates to the listener, and to the linguistic content of the message, rather than how the speaker 'feels' at any given time. But as we shall see below, it is necessary to parameterise all three forms of message content in order to adequately model the speaker's intent.

## 3. Speech Synthesis

Paralinguistic information, signalled by tone-of-voice, and speaking style, 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 normally use a different speaking style and tone-of-voice than when addressing a stranger or a wider audience. Speech synthesis must likewise become capable of expressing such differences.

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 [9], or in call-centre operations, then there is a clear need for the vocal expression of more than just the lexical and syntactic components of the utterance.

For unrestricted text-to-speech conversion, the problems of text disambiguation, and focus determination can be profound. They require a level of world-knowledge and discourse modeling that is beyond the capabilities of most text-to-speech systems. As a result, the prosody component of the synthesiser is often only provided with a basic or 'default' specification of the intentions of the speaker or of the underlying discourse-related meanings of the utterance, resulting in a flat rendering of the text into speech. This not a problem for some speech synthesis applications such as news-reading or information announcement, but if the synthesiser is to be used for interactive spoken dialogue, then the speech will be perceived as lacking illocutionary force, or worse, it will give the listener a false impression of the intentions underlying the utterance, leading to potential misunderstandings.

As part of the development of NATR, a communication aid [10], we designed an interface for the synthesis of conversational speech utterances, with specification of not just the lexical content of the desired utterance, but also for aspects of speaking style, including paralinguistic and extralinguistic features. We are testing this prototype with ALS patients, who need a speech synthesiser for essential daily communication with friends, family, and care providers, but we also envisage business uses of such a system for situations where overt speech may be difficult [11].

For example, a busy executive may want to telephone home to inform her partner that she will be returning later than usual because of a 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 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. The listener can read as much from the tone of voice in such cases as from the linguistic message.

## 4. Expressive Speech Processing

In December 1999, the Japan Science & Technology Agency issued a call for proposals under the CREST initiative 'Information Processing Technology for an Advanced Media Society', and a submission proposing the study of *"Expressive Speech Processing"* (ESP) was accepted the following April [12]. The acronym ESP is usually used to refer to Extra-Sensory Perception, but the coincidence is not accidental, as both processes purport to reveal a meaning hidden beneath the surface. The JST/CREST ESP Project started in July 2000 as joint research between ATR, NAIST Graduate Institute, and Kobe University, with contributions from ICP Grenoble, Keio and Chiba Universities, and from Omron Corporation's Verbal Interactive Technology project.

The goal of the five-year ESP Project is to produce a corpus of natural daily speech in order to design speech technology applications that are sensitive to the various ways in which people use changes in speaking style and voice quality to signal the intentions underlying each utterance, i.e., to add information to spoken utterances beyond that carried by the text or the words in the speech alone. The corpus is to include emotional speech, but also samples to illustrate *attitudinal* aspects of speech, such as politeness, hesitation, friendliness, anger, and social-distance. The most obvious applications of the resulting technology will be in speech synthesis, but the research also involves speech-recognition technology for the labeling and annotation of the speech databases, and the development of a grammar of spoken language in order to take into account supra-linguistic (i.e., paralinguistic and extralinguistic) information.

In order to provide speech data that are representative of the varieties of speaking styles found in a wide range of everyday situations, the speech should be that of ordinary people naturally expressing various attitudes and emotions in a variety of day-to-day interactive inter-personal situations.

When a corpus is based on read prompts (e.g., for the study of linguistic aspects of prosody, or for training HMM recognisers) the speakers' personal involvement is minimised by focussing their attention on the *form* of the text, and the resulting speech shows only the underlying syntactic and semantic relationships. Questions and statements, for example, don't originate from the speaker, but from the text, differentiated by the punctuation alone. The given/new relationships and focus information are similarly inferred, because the speaker is not the *originator*, but just an *interpreter* of the text.

In 'task-based' speech collection there may be more speaker involvement, but it is reduced to a paralinguistic minimum. The speaker is less motivated from internal desires than by a need to perform as required. Task-based elicitation produces speech with a prosody that signals not just the linguistic framework but also the pragmatic function, since, in a dialogue situation, the listener is as involved as the speaker. A request for information must be signaled as such, so as to to obtain a reply without any explicit scripting of the target speech. Task-based corpora are more natural-sounding, but are not in themselves natural. The speech may be unscripted, but the situation is contrived, and the speaker is *cooperating* rather than *operating*.

In all such cases, the need for a balanced scientific design frequently places unnatural requirements on a speech corpus, which render the content less than spontaneous. We can find many examples of such contrived-speech corpora in the literature

### 4.1 The Observer's Paradox

The term *Observer's Paradox* is attributed to Labov. It refers to a problem, faced by sociolinguists in particular, that, when observing or interviewing people to find out about their habits of speech, investigators will, by their own presence and participation, tend to influence the forms of speech that are used. In order to collect a corpus for the analysis of para-linguistic speech characteristics, we need observer-free recording. The corpus cannot be balanced or designed in the traditional scientific sense because our linguistic concepts may be biased by our views on the 'potential' of language use (e.g., Chomsky's 'competence') and influenced by the text-bound traditions of linguistic analysis and by existing corpora that are not representative of interactive daily conversational speech.

Control in corpus design is not the only cause of a lack of spontaneity. As we know from the Observer's Paradox, the presence of an observer can have an influence on that which is being observed. The presence of a microphone (or worse, of a recording engineer) can severely hamper the spontaneity of the speech. The alternative, of surreptitious recording, is ethically questionable (if not illegal) and results in data that cannot easily be shared or published.

In order to overcome this obstacle to natural data collection, we adopted what we term a 'Pirelli-calendar' approach for the ESP corpus [13,14]. In 1970 a team of photographers took 1000 rolls of 36-exposure film on location to an island in the Pacific in order to produce a calendar of twelve (glamour) images. We presume that the reason for this 3000:1 ratio of film to required photographs is that 'perfect' photographs cannot otherwise be guaranteed. We assume that if we can record an almost infinite amount of speech, and develop automatic techniques for processing it [15,16], to extract only the significant or interesting portions for further analysis, then we will be able to produce a corpus which is both truly representative and of sufficient coverage to allow us to define the full range of prosodic and speaking style variation and to formalise methods to describe its use in human communication.

### 4.2 Emotion in the ESP corpus

We have to date collected more than 250 hours of unconstrained spontaneous speech from a range of subjects using two collection paradigms. Both use high-quality head-mounted microphones for recording, but they differ in the recording medium; one using DAT, and the other MiniDisc [17]. The first (recorded onto DAT tape) is completely uncontrolled for content, with volunteers telephoning each other at regular intervals to talk freely for half-an-hour per session. Sessions were recorded at weekly intervals for a period of ten weeks [18]. The second (using the lighter and more portable MiniDisc recorders) employs volunteers who record their domestic and social spoken interactions for extended periods throughout each day. Analysis of these conversational-speech corpora in terms of emotion, using the broad-class labels *normal*, *happy*, *sad*, and *angry*, has proved extremely difficult, for two reasons.

First, the definition of *normal* appears to be highly context-dependent, as the speaking style varies according to both (a) familiarity with the interlocutor, and (b) type of conversation. Many of the extracts we examined (often just one side of a phone conversation with a friend) were textually very repetitive, but prosodically extremely rich, and they varied considerably in their functional meaning. Much of the language consisted of grunts and fillers, monosyllabic utterances or repeats of the same syllable many times, but expressing different meanings in each case.

Second, the *emotion* labels appear 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 so that we can distinguish between *revealed* and *intended* variants. For example, a schoolteacher controlling a class might not actually be angry when speaking in an angry manner to unruly students. The *speech* is angry, but the *speaker* is not. Conversely, the speaker might be feeling extremely angry, but manages for social reasons not to reveal it in the speech. A simple emotion label fails to differentiate between these cases, but most listeners can do so easily.

For the labeling of supra-linguistic variation in the speech database, each utterance must be evaluated separately for 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 moods and short-term emotions and the attitudinal states of the speaker, the pragmatic force behind the speech act, the voice-quality underlying the utterance (breathy, relaxed, pressed, forced), and so on. The list is not complete. The simplistic notion of a single switch for *emotion* in a paralinguistic speech synthesiser would appear to need considerable rethinking.

## 5. Annotating Expresive Speech

The ESP corpus is currently being labelled for speaking-style characteristics after being transcribed by hand and segmented automatically. The original intention was simply to label each utterance in the corpus in terms of four or five basic emotion categories but, as noted above, this proved to be extremely difficult.

In order to adequately describe the characteristics of the speech in the ESP corpus, we consider it necessary to distinguish at least 3 levels or categories of label, indicating speaker state, speaking-style, and voice-quality characteristics respectively. Labels are determined subjectively by an experienced labeller, after listening to the speech several times, to indicate how each section of the speech was perceived. Selection within the categories is by means of software offering pull-down menus (adapted from [19]) offering a limited range of choices for each category of label. The speech is categorised by the *combination* of the three label sets.

Since the goal of this work is two-fold; to provide a knowledge-base for research into speech and emotion (or *expressiveness*), and to provide a source database for expressive speech synthesis, our guidelines for the labelling are clear. The criterion in case of doubt is whether a given unit (a waveform segment) could be used in place of another given unit having the same set of labels in a concatenated synthesis utterance, without changing the perceived meaning or interpretation of the utterance. Meaning is here defined not just in terms of lexical and syntactic content, but also in terms of paralinguistic information and illocutionary force.

Examples of the descriptors used for each of the three categories are shown in Table 1. 'Speaker State' descriptors, are used to describe the speaker, rather than the speech, 'Speaking Style' descriptors, for labelling the way that person is talking, and 'Voice Characteristics' descriptors, to annotate the perceived acoustic nature of the speech.

The 6-point scales (Table 2) range from positive to negative in steps that are explained to the labellers as very noticeable, noticeable, only slightly noticeable. The lack of a *neutral* option forces the labellers to take a definite position rather than utilise an indeterminate label which is not well defined. Statistical methods are currently being applied to learn the mappings between measurable acoustic variation and the subjective category labels, using measures extracted from the speech signal.

| **STATE** | (*about the speaker*) |
|---|---|
| Purpose | a speech-act/CA label (open-class) |
| Emotion | happy/sad/angry/calm (4 classes) |
| Mood | worried/tense/frustrated/troubled/... |
| Interest | a 6-point scale from +3 to -3, omitting 0 |
| Confidence | a 6-point scale from +3 to -3, omitting 0 |
| | |
| **STYLE** | (*about the speech*) |
| Type | A speaking-style label (open-class) |
| Purpose | a speech-act label (closed-class) |
| Sincerity | insisting/telling/feeling/recalling/acting/... |
| Manne | polite/casual/blunt/sloppy/childish/sexy/... |
| Mood | happy/sad/confident/soft/aggressive/... |
| Bias | friendly/warm/jealous/flattering/aloof/... |
| | |
| **VOICE** | (*about the sound*) |
| Energy | a 6-point scale from +3 to -3, omitting 0 |
| Tension | a 6-point scale from +3 to -3, omitting 0 |
| Brightnes | a 6-point scale from +3 to -3, omitting 0 |

**Table 1. Three levels for describing supra-linguistics**

| | negative | positive |
|---|---|---|
| very noticeable | -3 | +3 |
| noticeable | -2 | +2 |
| only slightly noticeable | -1 | +1 |

**Table 2. Six-level forced-choice tendency scales**

### 5.1 Elements of SPEAKER STATE

The following categories are used to describe extra-linguistic or speaker-specific aspects of the spoken message. They refer to the state of the speaker as perceived from the wider context of the spoken signal. They do not require a knowledge of the speaker, nor of the context of the discourse, but a human annotator can infer much about speaker-listener relationships and the mental and physical state of the speaker from this level of information.
**Purpose** A conversation-analysis type label describing the pragmatic function of this section of the discourse. The labeller is free to suggest open-class labels to describe what the speaker is trying to achieve.
**Emotion** This category is deliberately constrained to the 4 emotions (happy / sad / angry / calm) that are offered by current emotion-enabled speech synthesisers.
**Mood** A label describing the speaker's mood (state of mind) using closed-class labels from a list including worried / tense / frustrated / troubled / etc. Used to complete the sentence: ``This person sounds ...''. (see also `mood' below).
**Interest** An estimate of the speaker's personal involvement in the discourse, marked on a 6-point scale.
**Confidence** A description of the speaker's personal confidence level marked on a 6-point scale.

## 5.2 Elements of SPEAKING STYLE

The following categories are used to describe para-linguistic aspects of the spoken message. They refer to the style of the speech as perceived from the limited context if a single utterance. They do not require a knowledge of the speaker, nor of the context in the discourse.

**Type** A unique descriptor decided by each labeller to associate a bundle of speaking-style labels (open-class) e.g., Angry1, Angry2, Greeting1, Bored3.

**Purpose** A closed-class speech-act label describing the illocutionary force of the utterance.

**Sincerity** A measure of the involvement of the speaker, i.e., the match between feeling and expression, expressed on a scale of insisting / telling / feeling / recalling / acting / reporting / citing/ etc., in the order of strong to weak involvement. The speech sincerity can differ with the strength of involvement as, for example, "I'm hot!" when used on a cold day (reporting/citing), or to inform the listener (insisting/telling), or to recall a feeling "I said "I'm hot", just like that, and he laughed" (acting/recalling).

**Manner** A description of the speaker's manner as understood from the speech sounds alone. As in 'This sentence was spoken .....ly'. This may be different from the attitude known to be held by the speaker from knowledge from a wider context of the discourse. Labels selected from a list including polite / rude / casual / blunt / sloppy / childish / sexy / etc.

**Mood** Used to complete the sentence: 'This speech sounds ...'. A description of the mood of the speaker as indicated by the sounds of the current utterance. This may be different from the mood of the speaker estimated from knowledge of the wider context. Labels selected from a list including happy / sad / confident / diffident / soft / aggressive / etc.

**Bias** An indication of the speaker-listener relationship as it can be distinguished from the speech. Labels selected from a list including friendly / warm / jealous / sarcastic / flattering / aloof / etc.

## 5.3 Elements of VOICE QUALITY

The following categories are used to describe physical aspects of the speech signal. They refer to the acoustic quality of the voice and can be marked on segments smaller than a single utterance. For practical purposes, a labeller-confidence assessment is also included at this level of labelling, but it is used to refer to all levels of signal labelling, rather than voice-quality alone.

**Tension** A subjective indication of 'strain' in the voice, measured on a 6-point scale.

**Brightness** A subjective indication of 'brightness' in the voice, measured on a 6-point scale.

**Energy** An indication of the range of variability and strength of vocal effort in the speech, measured on a 6-point scale.

## 5.4 Labeller CONFIDENCE

This measure is marked in order to allow the labellere to indicate how confident they feel in the choice of labels for each individual segment of speech. It is not directly related to voice quality, but is marked at the smallest unit size.

It will be noticed that there is some apparent duplication in the categories reported above. This is deliberate, and is needed to resolve those cases where the *speaker* attributes differ markedly from the *speech* attributes. For example, in the case of the schoolteacher referred to above, when the children hear the words "*be quiet!*" spoken in a certain tone of voice, they

will presumably obey, but on those rare occassions when the teacher is also actually angry, they will *instantly* obey, and in a different way. An element of fear will enter into the situation. People can hear the difference between simulated anger and the genuine emotion through perceptible differences in both the prosody and the voice quality of an utterance. Such differences must be annotated in our data so that the related differences in speaking style can then be analysed.

## 6. Discussion

Experience with labelling a subset of the corpus for speaking-style characteristics has led to the proposal that rather than label *emotion*, we should instead be considering *speaker-relationships*. The paralinguistic and extralinguistic cues in the speech reveal how the speaker relates to the *listener* ('friendliness'), and to the *content* of each utterance ('commitment'). We believe that these rather than emotion are the significant dimensions for a model of speaking style for the next generation of speech synthesis.

The proposed dimension of 'commitment', or *content-relationship*, governs the expressed or revealed sincerity of the speaker, including the expression of emotion, and revelation of any attitudinal biases. This dimension distinguishes the social roles that the speaker might be assuming from signs revealing the speaker's inherent attitudinal and emotional states.

The dimension of 'friendliness', or *listener-relationship*, governs the formality and the degrees of familiarity that can be expressed in the speech. The precise details are culture-specific and depend on inherent rank, age, sex, and familiarity differences, etc., but the speaker can manipulate this dimension freely within pre-determined limits on a case-by-case or a day-to-day basis.

## 7. Conclusion

This paper has identified a problem for speech information processing, and has presented a set of suggestions for its solution. The problem arises when supra-linguistic information is to be processed as part of the speech signal, as it must be if we are to process interactive conversational or daily speech. In order to parameterise the supra-linguistic information, we must first categorise it into extralinguistic parts that reveal speaker state information, and paralinguistic parts that reveal the speaker's intentions. Such information can be gained from an analysis of prosodic and voice-quality information in the speech signal.

A set of labels for describing these separate components has been proposed, and an interface for making use of para-linguistic information in speech synthesis has been described. Work is in progress to map between acoustic features of the speech signal and the labels determined from subjective evaluation. This work is still experimental, and these can only be described as partial results, but by sharing our experiences with the wider community, we hope that a compact and easily annotatable set of expressive speech tags can be agreed upon, and which will facilitate rapid specification of the supra-linguistic information carried in the speech signal.

Finally, the paper has argued that while the needs of industry and market-forces must also be taken into consideration in the design of resources for speech technology, there is also a need for bottom-up data-based research using corpora free from the preconditions imposed by task-oriented designers. The collection of such data poses considerable problems, but the approach described above has been shown to yield speech data of considerable interest and value.

**Acknowledgements**
The author wishes to express gratitude to all members of the JST/CREST ESP project, especially those working at ATR for their contributions to this paper, with special mention for Kimura Minako, Parham Mokhtari, and Carlos Ishii, and to Janet Cahn for her useful comments on an earlier draft of the paper. 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] ELRA: European Language Language Resources Association: web pages at www.elda.fr/catalog.html
[2] LDC: The Linguistic Data Consortium: web pages at www.ldc.upenn.edu
[3] Furui, S., Maekawa, K., and Isahara, H., "Spontaneous Speech Corpus and Processing Technology", in Proc SSST, 2002.
[4] 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.
[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 183-186, 1997.
[7] Sylvie Mozziconacci "Expression of emotion and attitude through temporal speech variations", in Proc Intl Conf on Spoken Language Processing, Beijing, 2000.
[8] Iida, A., Iga, S., Higuchi, F., Campbell, N., Yasumura, M., "A speech synthesis system with emotion for assisting communication", in Proc ISCA ITRW on Speech and Emotion, pp.167-172, 2000.
[9] Iida, A., Sakurada, Y., Campbell, N., Yasumura, M., "Communication aid for non-vocal people using corpus-based concatenative speech synthesis", Eurospeech 2001.
[10] Iida, A., Iga, S., Higuchi, F., Campbell, N., Yasumura, M., "Designing and Developing a Conversation Assistive System with Speech Synthesis and Emotional Speech Corpora", ISCA ITRW on Speech and Emotion, pp.167-172.
[11] Campbell, W. N., "What types of input will we need for expressive speech synthesis?" in Proc IEEE workshop on Speech Synthesis (CD-Rom), Santa Monica, 2002.
[12] JST/CREST Expressive Speech Processing project, introductory web pages at: www.isd.atr.co.jp/esp
[13] Campbell, W.N., "Databases of Emotional Speech", in Proc ISCA (International Speech Communication and Association) ITRW on Speech and Emotion, pp. 34-38, 2000.
[14] Campbell, W. N., "The Recording of Emotional speech; JST/CREST database research", in Proc LREC 2002.
[15] 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.
[16] Mokhtari, P, \& Campbell, W. N., "Automatic detection of acoustic centres of reliability for tagging paralinguistic information in expressive speech.", in Proc LREC 2002.
[17] Campbell, N \& Mokhtari, P., "DAT vs. Minidisc – Is MD recording quality good enough for prosodic analysis?", Proc ASJ Spring Meeting 2002, 1-P-27
[18] Ashimura, K., Campbell, N., "Telephone dialogue database of JST/CREST ESP project", in proc JSAI, 2002.
[19] Transcriber, public-domain speech annotation software: www.etcs.fr/CTA/gip/Projects/Transcriber