Towards Synthesising Expressive Speech; Designing and Collecting Expressive Speech Data

Nick Campbell

ATR Human Information Science Labs
Keihanna Science City, Kyoto, Japan, 619-0288

nick@atr.co.jp

Abstract

Corpus-based speech synthesis needs representative corpora of human speech if it is to meet the needs of everyday spoken interaction. This paper describes methods for recording such corpora, and details some difficulties (with their solutions) found in the use of spontaneous speech data for synthesis.

1. Introduction

Much of the research that has been carried out to date under the banner of 'corpus-based' speech synthesis has in reality been merely 'data-based'[1]. It is important for the "systematic study of authentic examples of language in use" [2] that the corpus materials form "a collection of naturally-occurring texts *(or speech samples, wnc)* chosen to characterise a state or variety of a language" [3], and that are "in themselves more-or-less representative of a language" [4]. However, the majority of databases used for speech synthesis research have been both purpose-designed and carefully produced, scripted and read, usually by professional speakers, under controlled or studio conditions. They illustrate 'language as system', but not 'language in use'.

Human speech is by definition always 'natural', but only by collecting speech from 'natural contexts' can we also be sure that the speech is representative of the variety of styles and manners that are likely to be encountered in everyday spoken interaction. This is of course not easy. Many past attempts at producing corpora of spontaneous speech for synthesis research have been limited by the acoustic, psychological, moral, and legal difficulties of capturing natural samples of ordinary speech in everyday interactive situations.

However, if speech synthesis is to evolve in the direction of spontaneity and expressiveness, in order to be used as a substitute for human speech in interactive situations, then there is a need for the corpora to illustrate *ALL* aspects of spoken interaction, including also the paralinguistic information on speaker states, attitudes, intentions, and relationships, upon which we can base our research. Several ways have recently been proposed to overcome the problems of such data collection.

2. Expressive Speech

The term 'spontaneous' is widely used to distinguish 'read-speech' or controlled 'laboratory' speech from what some consider to be more 'natural' speaking styles. However, even read-speech is, to some degree, spontaneous, and it would be exceptional if any two readings, even from the same person, were to be identical in all respects. We will instead use the term 'interactive' here to distinguish the expressive conversational forms of speech from the more controlled forms of speech that are common to many present-day corpora (or 'databases'). "Street-speech" [5] has been proposed as an equivalent term for representing the normal usages of the "man-(or person)-in-the-street", but it has other less welcome connotations. The difference which we wish to highlight is that although the degree of spontaneity in speech may be considered scalar and subject to personal interpretation, 'interactive' speech reveals much more than just the lexical and syntactic/semantic or structural content of the text alone, and that it cannot be adequately represented by a written transcription.

If a speech recogniser were to produce a perfect transcription of 'controlled' speech, then we would probably have enough information to be able to reproduce it adequately by means of speech synthesis. However, even perfect recogntion output from interactive speech would describe only its content, and give little or no information regarding its contexts. Most situated speech includes supralinguistic information, not just about the speaker (identity, age, sex, personality, mood, emotion, etc.), but also about the speaker's intentions; the degree of 'commitment' to each utterance, and the relationships (both long-term and short-term) with the listener. This information is suppressed or reduced in laboratory- or controlled- speech.

Interactive or 'expressive' speech is also marked by a high degree of non-verbal content. Speakers in conversational situations frequently use non-lexical 'noises' (grunts, groans, sighs, coughs, laughs, so-called "fillers", and backchannelling) and simple lexical items (such as "yeah", "uh-uh", "ummm", "ah", "hmmm", etc.) to express complex attitudes and intentions, and to indicate the intended interpretation of an utterance. These sounds are marked by prosodic and voice-quality variation to a much greater extent than are the lexical sequences of well-formed or content-information-bearing sentences, and their interpretation cannot easily be determined from the orthography alone.

3. Corpus-based Speech Synthesis

Recent trends in speech synthesis reflect a growing use of speech databases, not just for analysis or as training data, but also as a source of segments for concatenation. The Klatt synthesiser [6] encoded information about speech segments as acoustic feature-vectors describing discrete phoneme targets,

using sets of rules for interpolation between the phones and for their prosodic modification. Olive's later introduction of diphone segments [7] showed that there is also important information, encoded in the transitions between the phones, that cannot adequately be modelled by simple interpolations. Taking advantage of an increase in available storage memory and speed of computing, Sagisaka [8] extended the move away from acoustic modelling and towards retrieval of speech information by introducing large databases of isolated words and sentences as a source of units for concatenation. Campbell showed the influence of prosody on the spectral characteristics of the speech by using segments selected from databases of continuous fluent speech [9], resulting in the removal (or significant reduction) of subsequent signal processing of the output waveform, as can be seen in systems such as AT&T's NextGen, which is based on similar principles [10]. In retrospect, we can see that the trend in speech synthesis is to move away from explicit modelling of acoustics and towards methods that encode the speech information implicitly in the segments themselves, and then selecting among them by using models of the higher-level phonemic, prosodic and structural contextual features that account for the low-level acoustic variation. This has resulted in speech synthesis that models extra-linguistic features such as the age, sex, and personality of the speaker, as well as the linguistic content that can be defined by an orthographic transcription.

If we extend this trend towards true corpus-based speech synthesis, then we can anticipate that more and more information will be encoded directly in the speech signal, and that the art of synthesis will then lie in developing ways to access appropriate speech segments from features (or labels) that also describe the supra-segmental and paralinguistic events or states covered by the corpus. Synthesis will evolve beyond a modelling of the speech production mechanisms of 'language as system' to incorporate higher-level models of spoken-language communication, or 'language in use'. This will require finding methods for collecting speech data that are sufficiently varied and expressive, and developing methods for labelling the speech segments to describe the contexts for which they can be appropriately re-used in synthesis

Current speech databases, as evidenced by the latest research synthesisers (and those that are being used daily in commercial applications where they may go unnoticed), allow almost perfect resynthesis of speech in a given task domain and fixed speaking style, but in spite of their large size (many incorporating more than 30-hours of speech data) they are inadequate for the reproduction of different speaking styles. For many of the present-day applications this is not a serious problem, as an 'announcement' style of speech is suitable for such situations, but if speech synthesis is to become more widely used by the general public in everyday interactive or conversational situations, then there will be a need for more expressive styles of speech as output.

When a speech synthesiser is to be used in the place of a human voice, such as in prosthetic devices, games, database-interfaces, speech-to-speech translators, or robots, then the listener will expect more information from the speech than can currently be provided by existing technology. Paralinguistic information will be required, in addition to linguistic information, to signal the relationships between speaker and listener, and between speaker and speech-content. For example, few currently-available speech synthesisers (the HL-Syn research system [11] being a notable exception) yet control for voice quality in the pressed-to-breathy continuum, yet recent research [12,13] has shown that human speakers use significantly different laryngeal settings according to their relationship with the interlocutor. Pitch-range and speaking-rate also vary considerably more in interactive speech than in controlled speech.

4. Collecting Natural Speech Data

A recent government-funded project in Japan was set up to collect 1000 hours of spontaneous speech data (the Corpus of Spontaneous Japanese [14]) for speech-recognition research. Because the CSJ corpus was intended to stand in contrast to read speech, it was decided to record only conference presentations and studio monologues of prepared talks, typically of about 20-minutes in duration. The corpus is rich in the fillers, restarts, and disfluencies that are characteristic of spontaneous speech, but it also has a high word perplexity for Japanese (mean > 80, SD25 [15], in contrast to perplexities of 22 and 30 reported for the Japanese dialogues in the VERBMOBIL Database. [16]). This may reflect the formal nature of the speech content and the constrained speaking style (subjects were given 48-hours to prepare each monologue), which, while not exactly 'read', is at least well rehearsed, and only minimally interactive. The data is 'spontaneous' in the sense that the speaker is preparing each utterance in real-time, rather than reading them from a script, but the corpus does not represent many of the characteristics typical of conversational speech (i.e., of talking to achieve a response from the participating listener, rather than speaking with the goal of imparting new information). It presents many examples of 'language in use', but they are representative of only a limited subset of language-use in general.

In an effort to collect more conversational examples of language in use, so as to better understand the ranges within which speech can be used in interactive communication, another Japanese spontaneous speech corpus collection (the Expressive Speech Processing project [17]) has the goal of recording 1000 hours of natural spoken interactions for use in the development of an expressive speech synthesis system that will be capable of simulating 'interactive' speech.

When a person is speaking without being conscious of the speech process, a wider variety of controls may be used to express a greater range of information than that which can be obtained by the usual forms of elicitation and recording. However, it has been well known since Labov [19,20] that attempts to record 'natural' speech have to overcome the 'Observer's Paradox', i.e., speakers producing only what they expect the recorder wants to hear, or altering their speech in a way that they consider to be more socially acceptable.

It was therefore necessary to find ways to capture natural speech in-situ without the presence of an observer or obvious recorder having an effect on the speaking style and speech content. A large part of the preliminary research for the ESP project therefore concerned ways to collect speech data that were of high acoustic quality while being at the same time representative of spoken language in daily use [18].

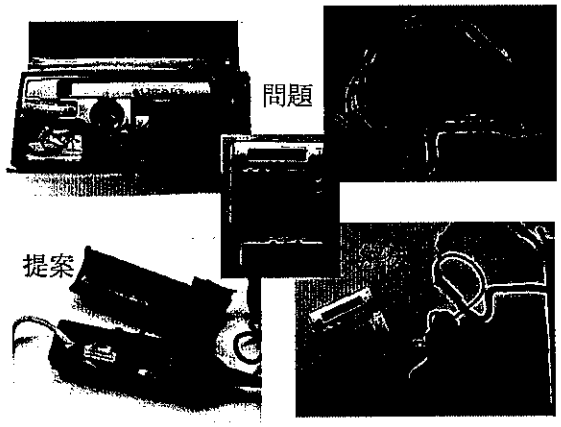

Figure 1. Adapting the mic power-supply for MD use. The lightweight studio microphone is fitted with a heavy and bulky three-pin adaptor. This also contains a power-supply which can, with some effort, be fitted into the extra battery compartment of the Minidisk recorder, resulting in a lightweight portable high-quality recording setup.

Two paradigms emerged; one, similar to Switchboard [21] or CALLLFRIEND [22], used telephone conversations, but with both speakers recorded locally onto DAT tape at 48kHz sampling rate using small head-mounted studio-quality microphones; the other made use of similar head-mounted microphones worn continuously throughout the day by volunteers, recording to small minidisk MD-Walkman recorders which were carried in a pocket or on a belt as they went about their daily routines. Adjustments were needed for the power supply to allow the microphone to be connected to the walkman (see figure 1), and tests were conducted to confirm that the differences resulting from compression of the speech between DAT recordings and MD were minimal [23].

For the telephone speech corpus, ten volunteers, balanced for age, sex, and mother-language, were recruited just to talk to each other, with no constraints on the content of the conversations other than time, for 30-minutes each week over a period of ten weeks. Pairing were balanced (see figure 2) to control for each of the above parameters, and familiarity (except with family members) assumed to increase throughout the collection period. Speakers were expected to have had no contact with each other outside the recording sessions.

For the everyday-speech corpus, there is only one side to any conversation, that of the microphone wearer, but the range of interlocutors varies greatly, including family members, friends, business acquaintances, shopkeepers, and virtual strangers, and the corpus has revealed a variety of voice qualities and speaking styles that was unanticipated, and which shows that conversational speech synthesis has many unexpected challenges yet to be met.

Most of the speech has now been transcribed orthographically, but as yet no phonemic alignment has been performed, due to difficulties arising from the spontaneous nature of the data. Part of the corpus has been labelled for speech-act, speaker-state, and speaking-style information [5], and work is currently underway to synthesise speech typical of each corpus.

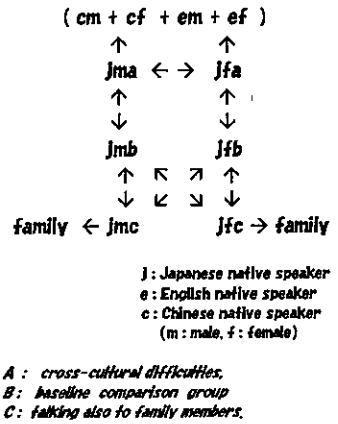

Figure 2. Speaker-pairings for the telephone conversations. Recordings were made of 30-minute conversations that took place each week over a period of 10-weeks. Speaker familiarity increased with each recording. No tasks or topics were set and the speakers were free to chat about any subject.

5. Synthesising Expressive Speech

From experience with CHATR [9] it was found that the best unit-selection target for concatenative speech synthesis is the speech waveform itself, or a parametric representation thereof. The quality of speech generated in this way is higher than that generated from prosodic and phonetic targets taken from natural speech, which is in turn better than speech generated from targets predicted from text. Of course, in situations where synthesis is to be used to generate speech, the synthesiser would be redundant if a waveform representing that speech were already available, but for the evaluation of various components of the synthesiser, this method has proved useful.

In the case of synthesising expressive speech, we can make further use of a waveform target for unit selection. By using a phonemically-balanced, controlled-speech database from the same speaker as a source of units for waveform concatenation we can create a target which has similar spectral characteristics but which is deficient in prosodic and voice-quality variation. Since these speaking-style characteristics have a smaller effect on the speech spectrum than do variations in phonetic characteristics, a same-speaker generated waveform is being used as a target for a second stage of unit selection [24]. Constraints (or voice-quality filters) are used to select further from among the various candidate segments thus produced.

This method is still experimental, and much work still needs to be performed for labelling the speaking-style characteristics in the corpus so that the pre-selection stage can be made more efffective, and the selection of non-lexical items, so-called fillers, definitely needs a functional control parameter. There are many ways to make a 'grunt', and each can be perceived as having a particular paralinguistic effect. Listeners have many years of experience in hearing human speech (and non-speech) sounds, and expressive speech synthesis needs a fuller understanding of these supra-linguistic aspects of speech if it

is to make significant progress. However, by basing the synthesis on a very large corpus of interactive speech, we are now in a position to begin such work.

6. Discussion

It is commonly believed, and many of the papers presented at research workshops and conferences support the view, that most of the problems of speech synthesis have already been solved, leaving only the implementational details yet to be researched. Indeed, we now understand and can successfully model or replicate many of the mechanisms of human speech. However, talking is not restricted just to giving information; and conversational speech also includes wheedling, whining, pleading, moaning, praising, groaning, laughter, and tears. We can talk with a smiling voice, a harsh voice, an authoritative tone-of-voice, a sexy voice, and many more.

Very few speech synthesisers are equipped with such facilities, yet if we expect to use them in our everyday lives, and for more than the simple provision of verbal information, then we as a community should start to collect data from which we can analyse, model, and reproduce these social human characteristics, as an extension of the more formal language and speech modelling which now forms the large part of speech synthesis research.

It has long been a basic tenet of scientific research that data should be controlled, in order to allow balanced experiments, but in controlling the content, we destroy the very naturalness that we need to study. Instead of designing speech synthesis top-down, through control of the data upon which it is modelled, we should perhaps aim for a bottom-up approach, analysing the ways that ordinary people us speech in their daily lives, and reproducing the many different kinds of information that speech conveys.

7. Conclusion

To clarify the title of this paper, corpora should be 'collected' rather than 'controlled' if we are to obtain a body of spoken language that allows us to study the needs of future speech synthesisers. Speech corpora cannot be designed; they must be captured. The paper has presented a brief overview of trends in speech synthesis, and has suggested that if synthesis is to be used as more than just a reading machine, then it needs to be equipped with the ability to perform in the same way as humans do, and to express paralinguistic information as well as linguistic content. The task of expressing extra-linguistic information has perhaps been solved by concatenative systems, but much research is still needed before a synthesised voice can express the subtle information that we have become used to hearing. Perhaps it is not a coincidence that the first ISCA workshop on speech synthesis, in Autrans 1990, resulted in a book called "Talking Machines". That was just the beginning.

Acknowledgements

The work of the Expressive Speech Processing project is funded by the Japan Science & Technology Agency, as part of CREST Research into Technology for an Advanced Media Society, Project #131. The author is grateful to the Eurospeech Organizing Committee for agreeing to host this special session, and to the volunteer subjects who have contributed the recordings of their daily speech interactions so that we can begin to understand the varied scope of speech information in everyday communicative situations.

References

[1] Hirose Keikichi, personal communication. 15/11/2002,
[2] Sinclair, J., *Corpus,Concordance,Collocation,* OUP 1991.
[3] *The Oxford Companion to the English Language*, ed McArthur & McArthur, Oxford University Press, 1992.
[4] Crystal, D., *A Dictionary of Linguistics & Phonetics,* Blackwell (3rd edition) 1991.
[5] Campbell, W. N., "Towards a grammar of spoken language: incorporating paralinguistic information", pp. 676-676 in Proc ICSLP2002, Denver.
[6] Klatt, D. H., "Software for a cascade/parallel formant synthesizer",Journal of the Acoustical Society of America, vol 67, 1980, pp. 971995
[7] Olive, J. P., and Spickenagel, N., "Speech resynthesis from phoneme-related parameters", Journal of the Acoustical Society of America, vol 59, 1976, pp993-996.
[8] Sagisaka, Y., "Speech synthesis by rule using an optimal selection of nonuniform synthesis units", Proc. IEEE-ICASSP88, 679-682. 1998.
[9] Campbell, W. N. & Black, A. W., "CHATR a multi-lingual speech re-sequencing synthesis system", pp45-52, *Technical Report of IEICE* SP96-7, 1996.
[10] AT&T NextGen: www.research.att.com/projects/tts (note that in the 'credits' section on this page, there is no mention of "CHATR-inside", but they explicitly note that "people active in the field will be aware of who invented what" – and that "periodic memory refreshment is welcome": AT&T licensed CHATR technology in 1997.)
[11] Stevens, K.N. and C.A. Bickley, "Constraints among parameters simplify control of Klatt formant synthesizer", J. Phon., 19, 161-174. 1991. Proc ICPhS 2003 in press.
[12] Campbell, W. N., "Voice Quality – the 4th prosodic dimension", in Proc ICPhS 2003 in press.
[13] Classen, K. and Wokurek, W., "Voice quality and discourse structure", in Proc ICPhS 2003 in press.
[14] Maekawa, K., Koiso, H., Furui, S., & Isahara, H., "Spontaneous Speech Corpus of Japanese", pp 947-952, Proc LREC 2000, Athens, Greece.
[15] Kawahara, T., Nanjo, H., Shinozaki, T, and Furui, S., "Benchmark test for speech recognition using the Corpus of Spoken Japanese", pp.135-138 in Proc SSPR2003, Tokyo.
[16] Kurematsu et.al., "Verbmobil dialogues: multifaceted analysis", p712-715 in Proc ICSLP2000, Beijing.
[17] The JST/CREST Expressive Speech Processing project, introductory web pages at: www.isd.atr.co.jp/esp
[18] Campbell, W. N., "Recording Techniques for capturing natural everyday speech", in *Proc Language Resources and Evaluation Conference* (LREC-2002), Las Palmas.
[19] Labov, W., Yeager, M., & Steiner, R., "A quantitative study of sound change in progress", Philadelphia PA: U.S. Regional Survey, 1972.
[20] Wolfram, W., Schilling-Estes, *American English*, Malden MA: Blackwell, 1998.
[21] Switchboard telephone-speech db: www.ldc.upenn.edu.
[22] CALLFRIEND: telephone-speech db LDC Catalog, 2001.
[23] Campbell, W. N., "Recording and Storing of speech data", LREC 2002 Workshop.
[24] Campbell, N., & Mokhtari, P., "Using a non-spontaneous speech synthesizer as a driver for a spontaneous speech synthesizer", pp.239-242 in Proc SSPR2003