Mapping from read speech to real speech

Nick Campbell
ATR-ITL

Abstract

For interpreting telephony, it is likely that we will soon be required to synthesise a wide range of utterance types in order to closely reproduce the voice qualities and speaking style of the translated speech. This paper describes the analysis of some English speech databases currently being collected at ATR for the synthesis of natural-sounding speech in an extension of the ν-talk principle. The corpora vary in content and style, from formal readings of lists of isolated words, through news-readings, and task-oriented dialogues, to completely spontaneous, unstructured monologue. The main task confronting us in the processing of such materials for synthesis concerns the development of a common set of labels to encode the significant prosodic and segmental variety in such a way that a compact set of units is capable of reproducing all the meaningful variation in the speech with the minimum of acoustic post-processing. In particular, the labelling of spontaneous speech confronts us with phenomena that were not considered relevant when the corpora were limited to discrete utterances, and we are increasingly having to identify larger, higher-level units of discourse structure for the adequate encoding of phonetic differences.

1 Introduction

Although speech synthesis has been an important tool in phonetic research for several decades already, we have yet to hear synthetic speech that sounds natural. Whereas isolated vowels and some consonants can be very well replicated and, with careful hand-tuning, even whole utterances can be mimicked, I am aware of no speech-synthesis-by-rule system that I could yet mistake for a human voice. The reasons for this are partly practical - constrained by needs to work efficiently in small computers using limited memory space - but also due to limitations of a more serious type: despite decades of analysis, we are still unable to model the spectral and intonational features of speech well enough to reproduce them adequately.

The best rule-generated synthetic speech that can be heard today is concatenative, taking small segments from recorded sequences of real speech and joining them to form novel utterances, but in the process the recorded segments lose much of their original naturalness. I maintain that the reasons for this loss of quality are two-fold: i) that degradation results from the signal-processing required to encode the segments and modify their prosody, and ii) from constraints in the recording of the original speech sequences themselves. Almost all segments for concatenation are taken from recordings of read speech, and although they may be phonemically representative, they are prosodically constrained to be invariant. Thus, what they encode adequately models the configurations of the vocal tract for a given sound sequence but fails to model the dynamic characteristics of the speech. The resulting speech lacks spontaneity and naturalness.

In this paper, I argue an extreme case, and propose that we remove the artificial constraint of limited memory size from synthesis design, and explore the implications of this change. This is not an altogether unrealistic concept as, in marketing terms, we can just as easily conceive of synthesis as being a service, rather than a product, distributed from a large central processor instead of being locally computed. Releasing the constraint does, however have severe consequences, in that it replaces the need for *modelling* of speech with a need for *characterisation* (or labelling) instead. Thus this paper addresses the phonetics of spontaneous speech, and attempts to define the optimal factors for description of natural speech such that it can be emulated in a synthesis system.

Take for example the modelling and prediction of segmental durations: this is normally the result of determining through analysis of large speech corpora a number of significant factors that combine in certain possibly complex ways and correlate highly with consistent variation in abstract representations of articulatory events at various levels in a spoken utterance. There is not yet consensus on what level of speech representation it may be most appropriate to model (see *e.g.*, [24, 5]), but there is considerable agreement about which factors are significant, including phonemic identity, position in the syllable, position in the phrase, etc. In synthesis, it is normal to predict the individual segmental durations on the basis of these factors and then to stretch or squash speech segments to fit the predicted time frame [?]. Two types of distortion occur here: i) the prediction is rarely perfect, and ii) in the changing of a segment's duration, frames have to be inserted or removed artificially. If we had a hugely finite corpus of speech as a source of units for concatenative synthesis then, instead of disruptively warping a segment's duration, it would be possible to select an appropriately timed segment from amongst the available variants. Furthermore, if that corpus were adequately labelled in terms of the contributing factors (i.e., with phonemic, syllabic, phrasal, etc. labels), then it would no longer even be necessary to predict numerical durations at all; it would be sufficient to select a segment from a part of the speech that was sufficiently similar in terms of prosodic and segmental characteristics to the desired target context. The durations (and other relevant acoustic features) would be contextually appropriate and natural by default.

Given the ease of recording, and the abundance of recorded speech material available today, the remaining challenge is to label the speech according to a small but sufficiently descriptive set of features so that all relevant variations can be indexed and retrieved. This reduces to a definition of the *perceptually salient* characteristics of speech, which in turn allows us to use a large speech corpus instead of a huge one without loss of quality.

2 Labelling speech

Scientific analysis requires controls, but as Barry has pointed out [2], in the acquisition of speech recordings, these are too often controls on production, with not enough concern for communicative effect. In the recording of lab speech (or of speech units for synthesis), the listener is replaced by the microphone. and the speech is always production-based rather than listener-oriented. Since in its natural form, speech is inter-personal and often functionally goal-directed, then the materials we collect and analyse may not be representative of what people do when they speak normally. In the analysis of speech, we need to replace production controls with statistical controls, and study instead large representative corpora of spontaneously produced spoken material. Such corpora are now becoming widely available but the tools for their analysis were developed for a more restricted speaking style. To cope with extremely large volumes of speech, the processing must be automatic. This section describes some of the corpora we use for synthesis and our methods for labelling large speech corpora, and discusses the types of speech variation that we find to be perceptually salient.

2.1 Corpora for synthesis

The ATR ν-talk system for non-uniform-unit concatenative synthesis of Japanese [28, 23] was developed using 5000 single words of speech as source units and has been tested with an alternative database of 503 phonetically representative read sentences [14]. In converting this synthesiser to produce English speech, we replicated these corpora and added several more. By the same speaker, we have readings of 5000 English words, a subset of these words read one at a time to form meaningful sentences, the same sentences read continuously, and 20-minutes of spontaneous monologue. From another speaker we have a series of task-related dialogues, performed in a multi-modal environment, both with and without access to vision of the interlocutor's face [21]. We have also tested the synthesiser using a source corpus of forty-minutes of radio-news speech [20]. These corpora were variously labelled at different sites using different transcription conventions. To compare them we had to relabel all to a uniform style.

2.2 Segmental labelling

It is very time-consuming to label large speech corpora by hand, but this can be automated to a large extent by using speech recognition technology, so that manual intervention can be limited.

We trained hidden Markov models (HMMs) to recognise segments of speech corresponding to the phonetic labels in a machine-readable pronunciation dictionary and generated networks of possible pronunciations for each word string of every sentence in the corpora. In the case of labelling, unlike pure recognition, we have neither to guess the utterance from scratch nor to be robust against speaker differences. Indeed, we were able to use Baum-Welsh re-estimation [9] to model the HMMs closely on the database, and to use orthographic transcriptions to constrain the alignments with accuracy comparable to human transcription [29]. Lexical sub-entries (such as 'gonna' for 'going to') cover significantly different pronunciation variants. The manual stage consisted just of transcribing the speech at the word level and aligning the orthography to the waveform.

Kohler [16] (see also Coleman [8]) has argued that although the articulation of a given string of words varies considerably under different speaking styles according to a cognitively-based reduction coefficient, dependent on speech act type, a linear segmental representation of canonical citation forms accounts best for such phonological reorganisation of speech. He shows that although segments may be elided or deleted in the production of fluent speech, a non-segmental residue remains to colour the articulation of the remaining segments. Such a canonical representation is easily accessible from a machine-readable pronunciation dictionary.

As Hirschberg points out [12], the major differences between lab speech and spontaneous speech appear to be prosodic (concerning speaking rate and choice of intonation contour), but there are also segmental differences. She notes for example that some disfluencies in spontaneous speech are marked by characteristic phonetic effects, such as *interruption glottalisation*, which is acoustically distinct from articulatorily similar laryngealisation. These characteristics cannot be differentiated from the segmental transcription derived from HMMs alone, but because of their prosodic dependencies, can to a large extent be predicted from a description of the prosodic context.

The labelling provides access to phone-sized segments of the speech waveform from which we can extract prosodic information in order to predict the finer articulatory differences and encode phonation-style characteristics without the need for marking them explicitly.

2.3 Prosodic labelling

Since in read speech, boundaries and prominences appear to be the most basic elements marking prosodic structure, we first need to be able to identify and label speech segments that appropriately signal these events. For example, a given speech unit immediately before a phrase boundary is likely to be very different from an equivalent unit immediately after one; it may be considerably lengthened, its amplitude low and decaying, and it may exhibit vocal fry. In the case of the same unit in a nuclear accented syllable, there will be differences in spectral tilt resulting from different vocal effort in phonation [22, 10, 26, 7] and in supraglottal articulation [15, 17] from the local hyperarticulation. By labeling context in both dimensions (segmental and prosodic), we encode these characteristics as an integral part of the speech unit.

Fundamental frequency, spectral tilt, energy, and duration are physical representations of suprasegmental characteristics, but they are secondary features that can be predicted from a higher-level representation of the utterance. They depend particularly on the context of a segment with respect to prominences and boundaries in the speech [6]. In order to label portions of speech according to differences in phonation type, we can therefore suffice by recognising segmental (*e.g.*, triphone) context as modified by local and global prosodic environments. The latter are derivable from prosodic contours across current, previous and following syllables, subject to modulation according to overall speaking rate, pitch range, etc.

The BU Radio News corpus [20] has been prosodically labelled according to the ToBI conventions [25] to differentiate high or low tones on prominent syllables and at intonational boundaries, and to mark the degree of prosodic discontinuity at junctions between each pair of words. Using this as training data, Wightman & Campbell [30] defined a series of acoustic, lexical, and segmental features derivable from the phone labels, the dictionary, and the speech waveform, that achieved detection of prominences at 86%, detection of intonation boundaries at 83%, and correct estimation of break indices (±1) at 88%. This was performed using a hybrid combination of a tree quantiser with Viterbi post-processing to maximise the output likelihoods, operating directly on the aligner output.

The acoustic features we extracted from the speech waveform for the autolabelling of prosody include (in order of predictive strength) silence duration, duration of the syllable rhyme, the maximum pitch target[1], the mean pitch of the word, intensity at the fundamental, and spectral tilt (harmonic ratio). Non-acoustic features considered are end-of-word status, polysyllabicity, lexical stressability, position of the syllable in the word, and word-class (function or content). These latter are all derivable directly from the dictionary used in the aligning.

Knowing if a segment is syllable-initial or final, and whether that syllable is prominent, phrase-final, or both, we are able to predict much about its lengthening characteristics, its energy profile, its manner of phonation, and whether it will elide, assimilate, or remain robust,

Re-synthesis tests, iteratively removing sentences from the radio-news corpus and synthesising them by concatenation of segments selected from the remaining utterances according to the prosodic criteria,

[1] Pitch targets are calculated using Daniel Hirst's quadratic spline smoothing to estimate the underlying contour from the actual f0.[13]

showed that much of the spectral variation in the segments is adequately coded in this way [4, 7]. In fact, when selecting segments across phrase boundaries, because the prosodic environment is specified, units are selected from pre- and post-pausal environments such that the 'silence' between them also includes an appropriate sharp intake of breath, which makes the resulting synthesis sound even more 'natural'[2]. It is not necessary to specify or model fine phonetic features explicitly if the higher-level description suffices to include them implicitly.

3 Mapping to spontaneous speech

Extending these labelling techniques to both the dialogue corpus and the spontaneous speech corpus, we immediately become aware of the need for an extra level of information to describe the higher-level structuring of the discourse and to indicate switches in style [2] that affect the global prosodic characteristics. Whereas the read speech was predictable in its rhythm and pauses, the unplanned speech exhibited bursts of faster and slower sections where the speaker varies her role, and much greater variation in f0 range and pausing as she expresses different degrees of confidence, hesitation, and uncertainty.

Transcribing the orthography now requires more than the skills of an audio-typist, and many decisions have to be made about conventions for marking disfluencies and restarts. We adopted the recommendations of Nakatani and Shriberg [19] for extending the miscellaneous tier to bracket interruptions in the speech flow, and added IFT (illocutionary force type) speech-act labels (after [27]). The following set was used:

> inform expressive good-wishes-response apology-response invite vocative suggest instruct promise yn-question confirmation do-you-understand-question wh-question action-request permission-request acknowledge yes no thank thanks-response offer offer-follow-up greet farewell good-wishes apology alert temporize generic-disfluency repair/restart hesitation phonetic-error laugh cough breath silence

As an example, the word 'okay' was said 140 times by one speaker in the dialogue corpus. It was variously labelled as acknowledge, confirmation, accept, offer-follow-up, and do-you-understand-question, etc.; twelve categories in all. The intonation, duration, and articulation varied considerably; sometimes short, sharp, and rising, on a high tone, sometimes slow and drawn out on a falling tone. Since we were able to find significant correlations between the intonation and the label for most of these cases (see [3] for a discussion), we continue in our assumption that instead of trying to predict and model the acoustic variations, we should instead be accessing them through higher-level labels. The appropriate labels for describing spontaneous speech must include speech act and discourse-level features if we are to capture sufficient variation.

Spontaneous speech appears to be most marked in terms of its rhythmic structuring, exhibiting greater ranges of variation with corresponding differences in phonation style. To appropriately describe these, we need perhaps to also develop a measure of the speaker's commitment to her utterance. Impressionistic comments such as 'she's thinking ahead', 'her mind's not on what she's saying', 'she's said this many times before', and 'she doesn't quite know how to put this' are triggered by differences in speaking style, but none of the labels we have considered so far are sufficient to mark such differences.

4 Speech style simulation

To date, synthesisers have not had to emulate spontaneous speech phenomena, and since Barnwell have been considered primarily as reading or announcing machines [1]. Perhaps they will never have to produce disfluencies or fragments, but we are already encountering computer speech in interpreting telecommunications, where the machine interprets the speech of a human in a dialogue and must faithfully translate not only the content of an utterance, but also its emotional colouring. To do this, we must be able to recognise and reproduce a variety of speech styles and speaker characteristics.

Given the assumption of unlimited memory and storage, the modelling of a different speaker simply requires switching to a different source database. By specifying characteristics of the speech in non-numerical form, any differences in *e.g.*, speaker-specific pitch range or speaking rate can be ignored - and the individual ways of expressing prominence or marking boundaries will be preserved.

[2] It should be mentioned here though that because of the limited size of this source database, simple concatenation of these selected units produces noisy synthetic speech, and some (distorting) signal processing is still required to reduce the discontinuities between the selected units.

However, modelling of a different speaking style requires recognition of the prosodic clues to that style in the input speech (feature extraction and labelling) and the selection of speech segments that are appropriate in phonation type from previously labelled data. The main tasks of synthesis thus depend on the efficiency of the labelling and the indexing of the speech. If we are to handle large corpora, this labelling needs to be done automatically; if they are spontaneous, then we need to be able to extend the current technology to cope with fragments, but recognising these from the acoustics alone is still an unsolved problem.

5 Summary

To summarise the main points of this paper, I have argued that for the efficient modelling of speech sounds (at least in the context of concatenative speech synthesis), it is not necessary to label fine articulatory details, nor to attempt the numerical prediction of prosodic attributes, but instead to use a higher-level specification of the environment in which they occur. We used a single limited set of phone labels to transcribe all the speech of one language (including its dialectical variants) but in order to encode the finer details of phonation style, have to complement this with a definition of the local and global prosodic contours. These contours are not specified directly, as this would lose us the ability to generalise from one speaker to another, but are specified in terms of the cognitive and communicative events which underlie them. The phone labels may be sufficient to encode the explicit content of the message of the utterance (as they map directly onto the orthography) but, and especially in the case of spontaneous speech, a significant part of the message lies in *interpretation* of *how* it was said. To encode that level of information, we need to incorporate labels for discourse and communication strategy. For read speech, it is probably sufficient to specify the prosody in terms of syntactic and semantic information alone, but for real speech we need also to estimate the state of mind of the speaker, her commitment to the utterance, and the role of that utterance in a greater discourse.

References

[1] Barnwell, T. P., "An algorithm for segmental durations in a reading machine context". Technical Report 479, MIT, Research Laboratory of Electronics. 1971.

[2] Barry, W. J., "Phonetics and phonology in speaking styles". In *Symposium on speaking styles, Proc ICPhS*, Stockholm, Sweden. 1995.

[3] Black, A. W., & Campbell, W. N., "Predicting the intonation of discourse segments from examples in dialogue speech". In *Proc. ESCA Workshop on Spoken Dialogue*, Hanstholm, Denmark, 1995.

[4] Campbell, W. N., "Prosody and the selection of source units for concatenative synthesis". In *Proc. ESCA Workshop on Speech Synthesis*, Mohonk, NY. 1994.

[5] Campbell, W. N., "Syllable-based segmental duration". In G. Bailly, C. Benoit, and T.R. Sawallis, editors, *Talking Machines: Theories, Models, and Designs*, pp 211–224. Elsevier. 1992.

[6] Campbell, W. N., "Automatic detection of prosodic boundaries in speech". *Speech Communication 13*, pp 343-354. 1993.

[7] Campbell, W. N., & Beckman, M. "Stress, Loudness, and Spectral Tilt", 3-4-3 in *Proc Acoustical Soc. Japan*, Spring meeting. 1995.

[8] Coleman, J. C., "The phonetic interpretation of headed phonological structures containing overlapping constituents". *Phonetics Yearbook 9*, pp 1-44. 1992.

[9] Entropic Research Laboratory, Inc, *HTK - Hidden Markov Model Toolkit* 600 Pennsylvania Avenue, Washington DC 20003.

[10] Gauffin, J., & Sundberg, J. "Spectral correlates of glottal voice source waveform characteristics", *Journal of Speech and Hearing Research 32*, pp 556-565. 1989.

[11] Hirschberg, J., "Using discourse content to guide pitch accent decisions in synthetic speech'. In G. Bailly and C. Benoit, ed, *Talking Machines*, pp 367–376. North-Holland, 1992.

[12] Hirschberg J., "Acoustic and prosodic cues to speaking style in spontaneous and read speech". In *Symposium on speaking styles, Proc ICPhS*, Stockholm, Sweden. 1995.

[13] Hirst, D., "Automatic modelling of fundamental frequency using a quadratic spline function" In *Travaux de l'Institut de Phonétique 15*, Aix en Provence, pp 71-85. 1980.

[14] N. Iwahashi, N. Kaiki, and Y. Sagisaka "Speech segment selection for concatenative synthesis based on spectral distortion minimisation". *Trans. IEICE vol. E76-A, 11*, November 1993.

[15] de Jong, K., "The supraglottal articulation of prominence in English: linguistic stress as localised hyper-articulation". in *Journal of the Acoustical Society of America 97(1)*, pp 491-504. 1995.

[16] Kohler, K, "Articulatory reduction in different speaking styles". In *Symposium on speaking styles, Proc ICPhS*, Stockholm, Sweden. 1995.

[17] Lindblom, B. E. F., "Explaining phonetic variation: A sketch of the H&H theory". In *Speech Producstion and Speech Modelling, NATO-ASI Series D: Behavioural and Social Sciences*, edited by H. J. Hardcastle and A. Marchal (Kluwer, Dordrecht), Vol 55. 1990.

[18] Moulines, E. & Charpentier, F., "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones". *Speech Communication, Vol 9, nos 5/6*, pp 453-467.

[19] Nakatani, C., & Shriberg, L., "Draft proposal for labelling disfluencies in ToBI". paper presented at 3rd ToBI labelling workshop, Ohio. 1993.

[20] Ostendorf, M., Price, P., & Shattuck-Hufnagel, S., *The Boston University Radio News Corpus*, forthcoming

[21] Park, Y. D., Loken-Kim, K., Yato, F., & Fais, L., "Analysis of the telephone and multi-media/multi-modal interface in an interpreting dialogue: Linguistic and paralinguistic behaviours". In *Proc JW-MMC 94*. 1994.

[22] Pierrehumbert, J. % Talkin, D. "Lenition of /h/ and glottal stop". In Papers in Laboratory Phonology II, eds. G. J. Docherty & D. R. Ladd, Cambridge University Press. 1992

[23] Sagisaka, Y., Kaiki, N., Iwahashi, N., and Mimura, K., ATR – ν-TALK speech synthesis system. in *Proceedings of ICSLP 92*, volume 1, pp 483–486. 1992.

[24] van Santen, J. P. H., "Some observations on the role of syllables and segments in speech timing". In *Proc. ATR Workshop on computational modelling of the prosody of spontaneous speech*, Kyoto, Japan. 1995.

[25] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J., "ToBI: a standard for labelling English prosody". In *Proceedings of ICSLP92*, volume 2, pp 867–870. 1992.

[26] Sluijter, A. M. C., & van Heuven, V. J., "Perceptual cues of linguistic stress: intensity revisited", In *Proc. ESCA Prosody Workshop*, pp 246-249, Lund 1993.

[27] Stenström, A., *An Introduction to Spoken Interaction*. Longman, London. 1994.

[28] Takeda, K., Abe, K., and Sagisaka, Y., "On the basic scheme and algorithms in non-uniform unit speech synthesis", In G. Bailly, C. Benoit, and T.R. Sawallis, editors, *Talking Machines: Theories, Models, and Designs*, pp 93-106. Elsevier. 1992.

[29] Talkin, D., & Wightman, C. W., "The Aligner: text-to-speech alignment using Markov models and a pronunciation dictionary". In *Proc. ESCA Workshop on Speech Synthesis*, pp 89-92, Mohonk, NY. 1994.

[30] Wightman, C., W., & Campbell, W., N., "Improved labelling of prosodic structures", IEEE Trans. Sp. & Audio, submitted.