Multi-Level Labelling of Speech for Synthesis

Nick Campbell & Akemi Iida†
ATR Interpreting Telecommunications Research Labs.
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, JAPAN
†*Keio University, Graduate School of Media and Governance*
5322 Endo, Fujisawa, Kanagawa, 252-8520 JAPAN
nick@itl.atr.co.jp, akeiida@sfc.keio.ac.jp, www.itl.atr.co.jp/chatr

ABSTRACT

Phonetics and prosody have traditionally been considered as discrete disciplines, having little overlap, yet both are considered to be essential and complementary components for speech synthesis. This paper describes an approach to modelling the variation in speech, based on a gestalt view, wherein the prosodic and phonetic aspects of each speech segment are combined. Phonation style is proposed as the third dimension needed to characterise the speech sounds for synthesis by concatenative of raw waveform segments. We show that by labelling a large speech corpus with such high-level features, much of the redundancy of information in the speech signal can be preserved, and the resulting speech output maintains variation under natural constraints.

1 INTRODUCTION

Since its beginning, speech synthesis has been used as a tool for phonetic research [1]. In this paper, we discuss the definition of speech information that would be minimally sufficient for controlling high-quality raw-waveform speech synthesis. In the hypothetical case, assuming unlimited computer storage and instantaneous search-and-retrieval of waveform segments, novel utterances could be generated by a simple re-ordering of real speech segments stored in memory and indexed according to a set of defining characteristics. If the source corpus was sufficiently rich, and the segments were appropriately selected, then any desired meaning and interpretation should be reproduced in a way *completely indistinguishable from real speech.* The aim of the current paper is to clarify characteristic features of natural speech such that an efficient index can be prepared for the retrieval of adequate waveform segments from such a corpus.

Previous research with the CHATR synthesis system [2, 3, 4] has shown that much of the meaningful variation in speech can be reproduced by concatenating phone-sized segments selected according to suitability criteria sensitive to their prosodic and phonetic environments [5]. Labelling of the speech data is performed at the canonical or broad-phonemic level, as predicted by dictionary lookup, with allophonic variation and fine phonetic detail preserved in the speech as a consequence of retrieval based on higher-level prosodic and phonemic contexts.

By selecting candidate segments from a speech corpus according to different levels of a) input information and b) labelling precision, CHATR allows testing of theories of speech description. For concatenative synthesis using raw unmodified waveform segments (which is a particularly stringent test), the efficiency of labelling is critical to the production of natural-sounding utterances. If the level of labelling is too finely detailed, then the system must able to predict the same fine level of detail and can be prone to error. However, even with broad labelling, good-quality synthesis can result from the simpler representation of the desired targets *iff* prosodic environments are respected. Broad labelling defers to the hierarchical dependencies in natural speech to achieve a naturalness of variation beyond that which can be statistically predicted.

For a single speaking style, it appears sufficient to index the corpus at the simple phonemic level provided that information about tonal and prosodic-boundary contexts is also available [5]. However, for different speaking styles, and particularly for expressing different emotions, phonemic-prosodic labelling may be insufficient. We recorded three one-hour corpora of emotional speech (angry, sad, happy) and used each for the synthesis of neutral, emotionally unmarked sentences having the same phone sequence and target prosodic characteristics. The underlying emotion in the speech is clearly perceived in the majority of the test sentences, suggesting that the voice-quality (or phonation style) may be different for each.

2 SPEECH INFORMATION

Speech carries many levels of information; not just about the text of the message, but also about the intended interpretation of its component words, the identity and mood of the speaker, and the urgency of the message.

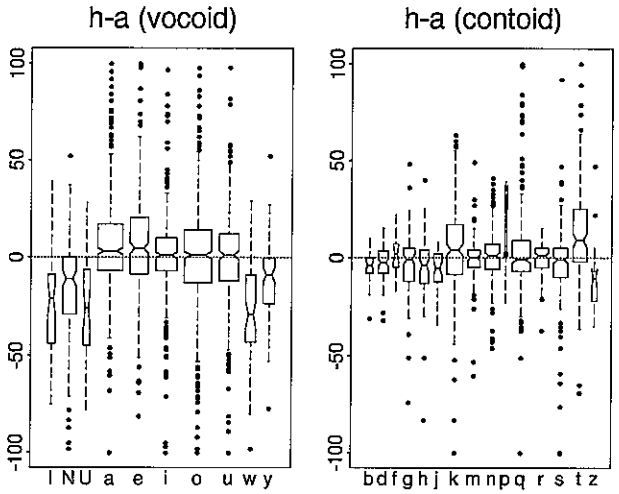

Figure 1: Duration differences (hand-labelled - auto-labelled) shown per phone type in milliseconds for a 30,000-phone database of Japanese. We can see that simple vowels tend to be longer in the hand-labelled data but that unvoiced (/I/ and /U/) and nasalised vowels are significantly shorter. Consonants show little change, but plosive closures are longer in the hand-labelled data. These differences probably result from different criteria for labelling transitional segments at the edges of vowels and in unvoiced sequences. (See also Tables 1 & 2.)

2.1 Text

Probably the most obvious meaningful information carried by speech is the text or *content* of the message ('Yes.' is not the same as 'No.'). The number of phone types used to label such lexical information determines the granularity of the labelling; diphthongs, for example, can be thought of as either one or two speech segments, affricates, long vowels and geminates likewise. On the principle of maximising tokens and minimising types, CHATR tends towards using the most basic label-set, and relying instead on natural transitions in the speech for modelling the fine context-specific variation. This allows for phone-based construction of novel words for synthesis, but requires strong distance measures and context description in order to ensure acoustic similarities and spectral continuity.

2.1.1 Is phonemic labelling adequate?

To differentiate between articulations of the same phoneme sequence (e.g., great eyes vs. grey ties) we do not require explicit allophonic or narrow-phonetic labelling (e.g., to mark the ±aspiration on the intersyllabic /t/), but use instead the prosodic information about the degree of stress on each syllable. The lower-level (allophonic) acoustic differences vary as a direct consequence of the different prosodic and phonemic environments and can therefore be captured implicitly by labelling the prosodic context.

2.2 Meaning

'Yes.' may not be the same as 'No.', but 'Yes.' may *mean* 'No.' if said, for example, with a with a hesitant rise-fall-rise tone rather than a simple falling one. The latter would imply acceptance whereas the former is more likely to indicate reluctance or doubt, especially when produced at a locally lower speaking rate. To reproduce such para-linguistic functions of ambiguous speech effectively, we need to model *all* the available acoustic cues to meaning, because when the prosodic information does not accord with the textual content, the natural redundancy of information in the speech signal is greatly reduced.

To convey prosodic differences in speech synthesis, signal processing has traditionally been used for modifying the timing and pitch characteristics of the speech, but at the expense of naturalness in the signal quality. A meaningful change in frequency is naturally accompanied by associated acoustic changes which arise from differences in vocal effort and formant excursions, etc. [6, 7]. These are rarely modelled accurately enough in the signal processing and the resulting speech lacks the richness required to signal subtleties of meaning.

Table 1: Percentile Duration Differences (in milliseconds) between auto-labelled and hand-corrected tokens for three speech segment classes from a forty-minute corpus of Japanese. (See also Figure 1)

	10%	25%	75%	90%
vowels	-28	-9	15	34
consonants	-19	-8	7	21
semivowels	-52.5	-34	-1	9.5

Table 2: Segmentation Error: differences (in milliseconds) between auto-labelled and hand-corrected boundaries in the speech signal for three speech segment classes from a forty-minute corpus of Japanese.

	mean (ms)	sd (ms)
vowels	18.77	19.84
consonants	13.35	16.86
semivowels	24.04	20.96

2.2.1 Is proso-phonemic labelling adequate?

By selecting speech segments according to criteria sensitive to both their prosodic and their phonemic environments, the above problem of signal incompatibilities is solved naturally. The essential decision about which prosodic events to label can be answered pragmatically. The prominences and boundaries in the speech, that signal the salience and segmentation of the message, are relatively easy to determine from the signal. We have found that by selecting speech segments that are appropriate in just the two dimensions of prosodic-boundary context and syllable-prominence context, a large part of their variance can be controlled.

Clearly, with a finite corpus of source units, there will rarely be perfect segments available to meet the desired prosodic and continuity targets for synthesis of novel utterances. However, because of the richness of the information in the raw speech signal, when there are enough 'close' segments that represent the desired speech characteristics, the Gestalt principle of 'closure' works to fill-in the gaps. The speech samples at [8](English), [9](German), and [10](Japanese) well illustrate the strengths and weaknesses of this approach. Particularly interesting in [10] is the mismatch between content and speaking style. The prosody is appropriate in terms of phrasing (although it sometimes deviates from the standard Tokyo accent) but the 'colouring' of the voice, particularly in the pre-pausal voicing, produces an impression of sadness that is inappropriate to the reading of a weather forecast. The selections were made from a small (one-hour) corpus of Japanese and demonstrate the continuity and naturalness of the synthesised speech, but they also show the limitations of selection according to phonemic and prosodic characteristics alone.

2.3 Emotion

The second author has prepared several speech databases for processing in CHATR, in order to determine the influence of emotional variation on segmental quality. For the work described here, she produced three one-hour read-speech corpora under different emotional states (happiness, anger, sadness), sustaining each emotion by involvement in the content of the texts, which were highly emotionally charged for happiness, anger, and sadness respectively.

Using the three corpora of emotionally-charged speech, we created source databases for synthesizing speech with emotion using CHATR. To test whether the emotion was perceptible in the speech even when producing sentences that are semantically neutral with respect to emotion, we synthesised five neutral sentences by selecting speech segments according to the same criteria from each source corpus. 18 university students were asked to identify the emotion types in the resulting fifteen utterances. Results showed joy: 51%, anger: 60%, sadness: 82% correctly recognized (see Figure 2). Chance results can be expected to be around 30%, so we conclude that the characteristics of the emotion are well preserved in the voice. The remaining

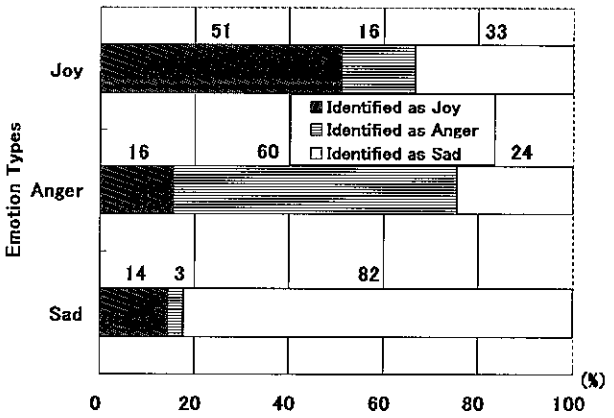

Figure 2: Identification rates for different emotions in speech synthesised from different databases

task is to identify the acoustic characteristics of the speech in each emotion type.

3 LABELLING SPEECH

Much of the information in speech can be labelled by automatic methods. This section describes some of the methods and compares differences between hand-labelled and machine-labelled corpora. It also discusses the limitations of automatic labelling and suggests some directions for future research.

3.1 Labelling phonemes

We employ automatic methods to produce an index of phonemic information time-aligned to the speech signal. Currently more than a hundred voices of Japanese have been labelled, and many have been hand-corrected. Results of a comparison between the hand and auto labelling methods reveal systematic but only small differences. Tables 1 and 2 show results from a 503-sentence corpus of Japanese, consisting of about forty-minutes of speech.

As Table 1 shows, 50% of the hand-corrected tokens were within -2ms and +9ms of the original auto-labelled boundary in the speech signal in the case of consonants, and within -2ms and +6ms in the case of vowels. 80%

Table 3: Results of automatic break-index labelling (summary of data presented at ICSLP-96).

	exact	within ±1	wrong
BX-2	1189	277	26
BX-3	713	283	106
BX-4	416	189	140

were within -14ms and +17ms (consonants) and -12ms and +32ms (vowels). The resulting segmental durations were within ±30ms for vowels and ±20ms for consonants, indicating a compensation between neighbouring phones where a boundary has been moved to lengthen one and shorten the other (see Figure 1). Typical examples of this are found where unvoiced plosives follow devoiced vowels, and at vowel-sonorant (/v/-/N/, /v/-/w/) boundaries. These results for Japanese speech compare favourably with those of [11] who performed a similar comparison for English.

3.2 Labelling prosody

By exhaustively predicting all the likely prosodic contours and dynamically aligning each to the observed contours of a given sentence we are able to label the phrasing of the utterance. The number of likely prosodic contours for a given word sequence is usually small. There may be more variety in interactive spontaneous speech, but the careful production of speech in the corpora that we are currently using for synthesis ensures that 'ungrammatical' renderings are unusual.

For Tokyo Japanese we were able to detect prominence correctly in 81% of the cases by differencing the fundamental frequency of the default and emphasised versions. Results of our automatic labelling of prosody were reported in [12]. For prosodic boundary labelling we employ break indices as in the ToBI system. Table 3 summarises results. Recent research results [13, 14, 15] show similar figures for other languages.

3.3 Labelling phonation

Although there are clearly prosodic cues to emotion (see Table 4) an ARX analysis of the source characteristics [16] of the speech waveforms also shows significant differences in the AV (amplitude of voicing) and GN (glottal noise) parameters between the different emotional speech styles. Distances of AV=41.9 and GN=23.1 are found between sad and happy speech, and of AV=26.6, GN=17.7 between sad and angry speech.

Table 4: Prosodic characteristics of human (top) and synthesised (bottom) emotional speech

HUMAN	F_0 (Hz)		Dur (ms)		RMS Pwr	
	Mean	SD	Mean	SD	Mean	SD
Joy	256.6	52.9	64.8	31.4	6.8	0.6
Anger	262.4	57.3	66.1	28.6	6.7	0.5
Sad	242.9	40.0	73.4	31.8	6.8	0.6
SYNTH	F_0 (Hz)		Dur (ms)		RMS Pwr	
	Mean	SD	Mean	SD	Mean	SD
Joy	243.8	38.8	83.2	84.7	6.2	1.2
Anger	256.2	41.0.	81.8	72.6	5.3	2.2
Sad	231.1	31.8	93.8	93.1	5.6	1.5

Work is in progress to use these differences in phonation style to also label emotional characteristics automatically. If successful, we could merge the three corpora and distinguish the different emotions using the third dimension of speech, rather than having to select from different databases as at present.

4 Conclusion

We have proposed, on the basis of tests using concatenative synthesis of raw waveforms, that for successful description of perceptually significant variation in the speech signal, the labelling should be performed at three levels: phonemic, prosodic, and phonatory. Two can be achieved automatically. We are exploring methods for the third.

References

[1] Fant, G. (1991) "What can basic research contribute to speech synthesis?", J. Phon. 19, 75-90.

[2] W. N. Campbell. Synthesis units for natural English speech. Technical Report SP 91-129, IEICE, 1992.

[3] http://www.itl.atr.co.jp/chatr.

[4] http://www.itl.atr.co.jp/chatr/interactive.

[5] W. N. Campbell and A. W. Black. Prosody and the selection of source units for concatenative synthesis. In J. van Santen, R. Sproat, J. Olive, and J. Hirshberg, editors, *Progress in Speech Synthesis*. Springer 1996.

[6] A. Sluijter & V. van Heuven, & J. J. A. Pacilly, "Spectral balance as a cue in the perception of linguistic stress", J. Acoust. Soc. Am. 101, 503-513, 1997.

[7] H. Traunmüller, Functions and limits of the F1:F0 covariation in speech, in PERILUS XIV, Department of Phonetics, pp.125-130, Stockholm University, 1991

[8] http://www.itl.atr.co.jp/chatr/j_tour/alda.html

[9] http://www.itl.atr.co.jp/chatr/j_tour/german.html

[10] http://www.itl.atr.co.jp/chatr/j_tour/fkt_tenki.html

[11] A. Ljolje, J. Hirschberg, & J. Santen, "Automatic speech segmentation for concatenative inverntory selection", Proc 2nd ESCA Synthesis Workshop, Mohonk, pp.93-96, 1994.

[12] W. N. Campbell "Autolabelling Japanese ToBI" Proc ICSLP-96 (Philadelphia) pp.2399-2402 (1996).

[13] A. Maghbouleh, "ToBI Accent Type Recognition", Proc. ICSLP-98, paper 632.

[14] "Automatic labelling of German Prosody", S.Rapp, Proc. ICSLP-98, paper 907.

[15] H. Vereecken, JP. Martens, C. Grover, J. Fackrell & B. Van Coile, "Automatic Prosodic Labeling of 6 Languages", Proc. ICSLP-98, paper 45.

[16] W. Ding, H. Kasuya, and S. Adachi, "Simultaneous estimation of vocal tract and voice source parameters based on an ARX model", *IEICE Trans. Inf. & Syst.*, Vol. E78-D, pp.738-743 (1995).