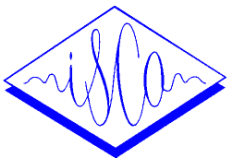

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

AUTOMATIC LABELLING OF VOICE-QUALITY IN SPEECH DATABASES FOR SYNTHESIS

Nick Campbell & Toru Marumoto

ATR Spoken Langguage Translation Research Labs.
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288 Japan
Nara Inststute of Science and Technology
e-mail: toru-m@is.aist-nara.ac.jp, nick@slt.atr.co.jp

ABSTRACT

A series of experiments was performed to determine the extent to which voice-quality differences could be labelled automatically in a speech database. Using speech corpora of three different speaking styles from the same speaker as test material, hidden-Markov models were trained to distinguish the prosodic and acoustic characteristics of each style, and were used to re-label the voiced-segments in order to provide a single, merged, labelled corpus. Perceptual tests of speech synthesised by concatenation using CHATR showed that both prosodic and voice-quality cues to stylistic variation (in this case emotion) can be detected and labelled by the trained models. However, speech synthesised from the original separate databases was perceived as being more expressive.

1. INTRODUCTION

Concatenative speech synthesis has shown the potential for producing natural-sounding speech, but when the component speech waveforms are joined without the use of signal processing, then there is need for a large, varied, and well-balanced database of source units if the quality of the speech prosody is also to be maintained.

When speakers read long sets of phonemically- and prosodically-balanced sentences for recording a source-unit databases they can become tired or frustrated and the diffferences in voice-quality between samples taken at the beginning of a session and those taken towards the end can be quite great.

If segments from these two extremes were to be concatenated in the same synthesised utterance, then the abrupt jump in voice-quality would become a source of noticeable degradation in synthesis quality.

This effect is particularly noticeable when speech is recorded over a period of several days or months, when other factors than tiredness can also influence the voice quality. It is therefore necessary to label voice-quality as well as prosody and phonemic characteristics when preparing a large spoken corpus as a source of synthesis units.

As a first step towards the automatic labelling of such voice-quality diffferences, we performed experiments on the detection and annotation of speaker-state using data from one-hour recordings of three different emotionally-charged passages read by the same speaker.

2. CONCATENATIVE SYNTHESIS

The CHATR concatenative speech synthesis system [1, 2, 3] produces human-sounding speech from a natural-speech waveform database in two stages. First, as an off-line one-time process, the phonemic, prosodic, and acoustic features of the original speech corpus are labelled and a segmental index is produced. Then, in real-time, a target specification is predicted from the input text to be synthesised, and a sequence of units for concatenation is selected from the database according to (a) the degree to which they fit the target specification, and (b) the degree to which they can be imperceptibly concatenated.

The resulting synthesised waveform has the acoustic characteristics of the reference speaker and the phonemic-sequence and prosodic characteritics appropriate for the text of the synthesised utterance. Previous work by Iida [4, 5] switching between the different source databases of emotional speech has shown that listeners can reliably discern the intended emotion from the synthesised speech according to the type of source database used.

It is of interest to us at this stage to know whether we can merge the three source databases into one large database and then access the different voice-qualities by labels according to characteristics which identify the various speaking styles. For this, we need an efficient method of labelling the voice-quality of the component segments.

3. MERGING THE SPEECH DATA

By merging the three databases into one, we increase the coverage and can better identify redundant or duplicated segments, which can then be pruned out in order to reduce the database size without loss of synthesis quality.

However, it may be mistaken to assume that all segments in each original database actually share the same emotional colouring, and rather than classify them according to source, we should determine the charactersitics of each segment individually.

In this way, we can ensure more efficient use of the individual segments, and can introduce an additional 'weak' category to indicate those segments which are not particularly marked for any emotional content. These latter may be used for the synthesis of any speaking style, and thus actually increase the

effective coverage of the three source-unit databases.

3.1. Prosodic differences

Prosodic characteristics of the three databases were estimated and found to be significantly different. Table 1 shows results of Student's t-test comparisons for the vowel /a/ under each pair of conditions.

The following parameters were determined to distinguish the three speech databases:

[angry]	pow	$(p(t) - p_{min}) \times 1.1 + p_{min}$
	dur	$dur \times 0.98$
	f0	$(f(t) - f_{min}) \times 1.125 + f_{min} + 7$
[happy]	dur	$dur \times 0.95$
	f0	$(f(t) - f_{min}) \times 0.625 + f_{min}$
[sad]	pow	$(p(t) - p_{min}) \times 0.95 + p_{min}$
	dur	$dur \times 1.08$
	f0	$(f(t) - f_{min}) \times 0.6875 + f_{min} - 10$

These modified values were tested as targets for selection of units from the combined database and proved to be effective in portraying the desired emotion, but it was noted from Iida's work that even speech with the same prosodic characteristics can convey different emotions if synthesised from the different source databases. This implies that voice-quality characteristics also signal affect, so in contrast to the above prosodic dimensions we attempted to distinguish the data also in terms of spectral characteristics for an improved labelling.

3.2. Spectral differences

Several measures have been proposed to estimate spectral characteristics of speech segments [6, 7] but for this test, we adapted the hidden-Markov models already being used for CHATR database segmentation and auto-alignment. We included spectral-tilt information based on the Harmonic and Amplitude measures proposed by Hansen and subsequently used by Sluijter [8, 9, 10] in her experiments on the voice-quality of spoken Dutch and English.

Hansen and Sluijter were primarily interested in the prosodic and linguistic aspects of voice quality, rather than in the detection of emotion, but since the breathiness in happy and sad speech can be readily distinguished from the more forceful quality of angry voice (for example) we considered it valid to use the same measures for this purpose. More importantly, these measures can be easily and automatically derived from the waveform spectral sections without the need for inverse filtering of the signal, which can be a very time-consuming task often requiring manual intervention.

Table 1 shows results of Student's t-tests for all tokens of the vowel /a/ in the source databases under each combination of the three speech conditions. It confirms that, with the sole exception of the first-formant bandwidth in one case (happy versus sad speech), the spectral-tilt parameters are significantly different for each.

Table 1: Results of t-tests for vowel /a/ under three emotional speech conditions, showing that the spectral-tilt parameters are significantly different for each. A, H, and S stand for angry, happy, and sad speech respectively. H1-3 are the frequencies of the first three harmonics of the spectrum, A1-3 are the first three formant amplitudes, and B1 is the bandwidth of the first formant.

	A vs H	H vs S	S vs A
H1-H2	-31.89	-7.77	38.25
H1-H3	-5.05	-2.89	7.30
H1-A1	-25.18	-9.94	33.13
H1-A2	-24.85	-6.87	29.90
H1-A3	-14.60	-12.69	25.27
B1	-16.46	-0.014	16.12
f0	5.15	13.88	-17.97
duration	10.0	-21.09	11.90
power	6.44	10.66	-15.43
df	10453	9986	9561

4. RE-LABELLING THE SPEECH

Three initially similar sets of phone-based hidden Markov models were re-trained on the three databases separately, using the following parameters: monophone single-gaussian 3-state HMM models with a frame rate of 10ms. Viterbi re-estimation was performed using prosodic and/or spectral vectors representing the speech signal, making use of the known phonemic transcriptions, to model the characteristics of the three databases individually.

Each HMM set was tuned to the specific characteristics of one database after initial estimation on smaller samples of the data. Although generalisation is necessary when training HMMs for speech recognition, there is no such requirement for speech synthesis labelling, where the data is closed and finite, and the need is for fine-feature detection rather than for open word recognition.

Input for the re-labelling was aligned speech vectors with their phonemic transcription; output was a set of likelihoods for each phone in the speech sequence as determined by the separately trained HMMs.

The different acoustic likelihoods produced by the database-specific models were taken as indicators of the most likely voice-quality type for each segment in the merged database. It was assumed that if the likelihood produced by one HMM was significantly higher than the other two, then the model which produced that score was better matched to the prosodic or acoustic characteristics of that segment of the speech.

If there was little difference between the two highest likelihoods, then it was assumed that the speech segment was weak in its characteristics and not particularly marked for any voice-quality type. Thus the three HMMs were used in combination to label four classes of voice quality. Only voiced segments of speech were re-labelled, since the spectral and

Table 2: The three methods of labelling give slightly different results, but all indicate a large number of 'weak' segments.

spectral	angry	happy	sad	weak
angry 19885	7340	5847	3291	3407
happy 20096	3941	9055	3331	3769
sad 15243	2490	6414	3607	2732
split [%]	24.9	38.6	18.5	18.0
prosodic	angry	happy	sad	weak
angry 18679	6169	4813	3966	4937
happy 19143	4899	4905	4946	5346
sad 14701	2362	3237	5723	3921
split [%]	24.3	23.5	26.5	25.7
both	angry	happy	sad	weak
angry 19885	7285	4492	4905	3203
happy 20096	4024	6282	5889	3901
sad 15243	2336	3432	7089	2386
split [%]	24.7	25.7	32.4	17.2

prosodic quality of unvoiced segments was assumed to provide less information for this preliminary stage of the work.

Re-labelling the segments resulted in different categorisations of segments according to each method. Prosodic information alone resulted in more 'weak' segments, evenly splitting the data into four categories (see Table 2). Spectral information, on the other hand, resulted in more segments being labelled as happy. Merging the two information sources resulted in a predominence of sad labels.

5. EXPERIMENTAL PROCEDURE

Since the 'right answer' is very difficult to determine except subjectively by intensive listening to hours of speech, we tested the efficiency of the re-labelling by perceptual evaluation of speech synthesised according to the new label information. There are three ways that source-unit segments can be selected for synthesis under the conditions of the present experiment. The first uses the current method of database independence; different emotions are indicated in the synthesised speech by selecting the segments from the appropriate (and separate) source database. The second is the proposed method of relabelling the segments according to emotion and selecting them from the single large merged database; in this case, their label may not correspond with that of the original source database from which they came. The third method is to use a merged database but to select according to prosodic targets rather than by segment labels. All three methods were tested.

This resulted in six synthesis methods for each of the three emotion types: (I) from re-labelling using spectral information alone, (II) from re-labelling using prosodic information alone, (III) from a combination of spectral and prosodic re-labelling, (IV) selection by prosodic target alone, (V) selection by prosodic target as well as re-labelling information, and (VI) selection from the original separate databases.

Since in four of the above synthesis methods the prosodic targets are held constant regardless of emotional type (except IV and V), any percepts of 'emotion' in the resulting synthesised speech must come from features inherent in the labelled information. Types I, II, and III allow an estimation of the contributions of voice-quality and prosody, both separately and in combination. If two utterances having almost the same prosodic and phonemic sequences can be distinguished (as in the case of type I), then we will be justified in taking voice-quality an contributing factor by which to select segments from a mixed-quality speech database, and can have some coinfidence in the proposed re-labelling process.

Speech waveforms were synthesised according to each of the above six methods for twelve test sentences. Semantically neutral sentences such as "Today is Tuesday", and "I didn't know she likes honey" (in Japanese) were synthesised according to each of the above methods. Listeners perceptions of the speaker's emotional state were used as judgements of the adequacy of the re-labelling. These were listened to in randomised order by twelve unpaid subjects on two occcassions. Subjects were unaware of the purpose of the experiment. The first session made use of a DAT tape recorder and headphones, presenting the samples at a fixed rate; the second used an internet web-page, allowing the respondents to adjust the pace of their listening. No major differences were determined as a result of test-type, and the results will be pooled in the following discussion.

6. RESULTS

The results confirm that correct detection of the intended emotion occurs at levels greater than chance in the majority of cases, but they do not show a clear difference between the methods. Only type VI, the original separate-database method, shows a strong distinction beteween the speaking styles. This can be taken to indicate that the speaking style information is shared between prosodic and acoustic aspects of the speech, but calls into question whether segments labelled as 'weak' are actually unmarked for emotion. The fact that type VI results are so much stronger indicates that all segments in each database may be characteristic of the given speaking style.

Table 3 shows the confusion scores from the perceptual tests. Chance score is 33.3%. Scores that are significantly above chance are marked with an asterisk in the table. Note that some scores are significantly well below chance, indicating that happy-sounding speech is harder to achieve by these methods (on semantically neutral sentences).

Comments from the subjects indicated that several of the sentences (being taken from news stories) were incompatible with a happy reading so further investigation is needed into confusions caused by textual interference.

The main confusion is between happy and sad speech, but the results also show that, in some cases, intended anger is perceived as joy. This would be unfortunate if the speech synthesis system were to

Table 3: Recognition rates (%) for the intended speaking style on semantically-neutral text. Chance detection is 33.3%. Confusions for the two competing responses are shown in the second row of each class. It is interesting to note that none of the methods reliably produced happy-sounding speech.

	type I	type II	type III	type IV	type V	type VI
angry	41.6 *	43.0 *	43.1	64.8 *	45.2	61.4 *
h/s	29.5/28.9	32.5/24.5	36.2/20.0	27.9/7.3	42.2/11.9	27.8/10.8
happy	23.0	11.2	23.1	10.0	25.7	58.2 *
a/s	30.2/46.8	28.9/59.9	25.9/50.9	48.8/41.3	36.8/37.5	22.0/19.9
sad	46.2 *	52.8 *	44.8 *	51.6 *	42.6	88.3*
a/h	35.0/18.9	29.2/18.1	35.9/19.3	43.3/5.1	36.0/21.3	7.0/4.7

be used as a communication aid, and is surprising as similar confusions do not occur often when listening to the original speech.

7. DISCUSSION

The results of the perceptual tests show that both forms of labelling are effective, and indicate that spectral information can be used in addition to prosodic information to achieve voice-quality labelling, but confirm that the use of separate databases is still to be preferred for the synthesis of controlled speaking styles.

Of particular interest is the relation between types I to III; which shows that spectral information or voice quality codes as much information about the state of the speaker as does the prosodic information. The interaction does not compound (type III is not significantly better) but there may be conflicting cues if both forms of information are not taken into account.

We conclude from these results that both spectral and prosodic information are necessary for labelling differences in speaking style, but further research is needed to determine whether every segment should be labelled for speaking style or whether there are particularly marked sections of the utterance for which extra care must be taken. It may be the case that the sentence-final segments alone carry enough information to bias the interpretation.

8. CONCLUSION

This paper has reported work extending that previously published in [11, 12]. We have tested a method for automatically identifying spectral and prosodic characteristics in the speech signal that correlate with perceived differences in the emotional state of the speaker. These were used to label the voice-quality of individual speech segments for concatenative speech synthesis. The resulting percept is not as clear as when the segments are taken from separate speech databases but a tendency to recognise the intended emotion was observed, even when labelling was performed from voice-quality alone. We conclude from this that similar methods might also detect the differences in speech quality that arise from speaker tiredness, and thus help to maintain a higher quality of concatenation when large quantities of speech data are collected over a long period of time.

REFERENCES

[1] W. N. Campbell: "Synthesis Units for Natural English Speech", Transactions of the Institute of Electronics, Information and Communication Engineers, SP 91-129, pp 55 - 62. 1992.

[2] W. N. Campbell, "CHATR: A High-Definition Speech Re-Sequencing System", Proc ASA/ASJ Joint meeting (Hawaii) 1996.

[3] W.N. Campbell, A.W. Black, IEICE TEch Rept. SP96-7 1996.

[4] Akemi Iida, Nick Campbell, Michiaki Yasumura "Emotional Speech as an Effective Interface for People with Special Needs" APCHI'98(Asia Pacific Computer Human Interaction 1998), 1998/7/15

[5] Akemi Iida, Nick Campbell, Michiaki Yasumura "Design and Evaluation of Synthesised Speech with Emotion"、 Journal of the IEICE, 40,2, Feb 1999.

[6] M. Jackson, P. Ladefoged, M. K. Huffman, & N. Antoñanzas-Barroso, "Measures of spectral tilt", UCLA Working Papers in Phonetics, 61, 72-8, 1985.

[7] Gauffin, J. & Sundberg, J. "Spectral correlates of glottal voice source waveform characteristics", pp 556-565, JSHR 32. 1989.

[8] Agaath Sluijter, "Phonetic Correlates of Stress and Accent", PhD Thesis, Holland Institute of Generative Linguistics ISBN 90-5569-013-9.

[9] A. Sluijter & V. van Heuven, "Spectral balance as an acoustic correlate of linguistic stress", J. Acoust. Soc. Am. 100, 2471-2485, 1996.

[10] 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.

[11] T. Marumoto and N. Campbell, "Speaking Style and Concatenative Speech Synthesis", in Proc ASJ Spring Meeting, 1999.

[12] T. Marumoto, "Voice-quality labelling", Master's Thesis, Graduate School of Information Science, Nara Institute of Science and Technology, 1998.