TR-IT-0137

An Investigation of the Quality of Concatenation of Speech Waveforms

Andrew J. Hunt
アンドリュー ハント

Alan W. Black
アラン W ブラック

1995.11

Three unit concatenative speech synthesis systems are currently being developed or supported by ATR Interpreting Telecommunications Research Laboratories: ATR ν-Talk, CHATR and SUBPHONET. An important requirement for the generation of high quality synthetic speech by these systems is that the joining of units produces smooth output. The research presented here aims to develop signal processing estimates of the quality of concatenation. A tightly controlled perceptual experiment was carried out to obtain subjective judgement of the quality of isolated words produced by concatenation of two units (i.e. with only one concatenation point). A range of standard signal processing measures was evaluated for the ability to predict the subjective judgements. These measures included power, fundamental frequency, cepstrum and MFCC, two compressed forms of MFCC, and two dynamic variants of MFCC. The experimental results show that MFCC parameters provide the best basis for predicting concatenation quality, and that a combination of power and MFCC can significantly improve upon the use of MFCC alone. Moreover, the compressed versions have similar predictive accuracy, a result which allows us to trade-off predictive accuracy and computational and storage requirements. In comparison to the improved cepstral representation used in ATR ν-Talk and SUBPHONET, a vector quantisation representation of MFCC has slightly better predictive accuracy but requires less than 1% of the space for storage.

1 Introduction

In the speech synthesis systems developed by ATR Interpreting Telecommunications Research Laboratories, ATR ν-Talk [1, 2] and CHATR [3, 4], speech is produced by concatenating units from natural speech. An important criteria for the output of natural speech is that the units being joined produce "smooth" speech. In other words, the systems must avoid selecting and joining units which introduce discontinuities into the output that degrade the overall speech quality.

Current and previous work on ATR ν-Talk and CHATR has employed the *cepstral distance* as an estimate the smoothness of concatenation of two units. The cepstral distance is derived as the Euclidean distance between cepstral vectors [5] close to the cut-points of two speech segments and is one estimate of spectral similarity. The assumption is that joins should be perceptually better when the spectrums of the units being joined are similar (i.e. the cepstral distance is low) and that they should degrade as the distance increases.

Both systems have used the *Improved Cepstrum* [6] in this calculation. More recently, the CHATR synthesis training process has used *Mel-Frequency Cepstral Coefficients* (MFCC) [7, 8] and the unit selection procedure has utilised a *Vector Quantisation* (VQ) estimate of MFCC. The spectral representations of these three signal processing methods are substantially different. However, the different measures have not been experimentally evaluated to determine their effectiveness in predicting the perceptual quality of speech synthesis unit joins. Moreover, these three spectral representations vary substantially in their storage requirement, in the CPU-time required to process them and in the CPU-time required to calculate them for large databases.

The research described in this technical report attempts to predict the quality of concatenation of units using signal processing measures. The distances provided by the three cepstral representations already in use were evaluated and several other candidate signal processing measures were investigated. The results show that very substantial reductions can be achieved in the storage space and processing requirements in comparison to previous methods, while simultaneously improving the prediction of concatenation quality. The results also show that the combination of cepstral distances and power differences improves upon this prediction. Finally, the speech data and perceptual results provided by this research remain available for future research of new perceptual prediction measures.

1.1 Overview

Section 2 describes the design and execution of the perception experiment in which isolated word speech tokens were produced by controlled concatenation and judged for quality by human subjects. Section 3 describes statistical tests which evaluate the perceptual results. Section 4 describes the results of statistical tests in which the perceptual quality of the speech tokens is predicted from a range of signal processing measures, including those discussed above. Sections 5 and 6 discuss the application of these results to speech synthesis and consider future work.

2 Design and Testing of Speech Tokens

2.1 Speech Data

A series of isolated word speech tokens was constructed by concatenation of parts of clearly articulated isolated words. The isolate words were selected for concatenation from the more than 5000 words in the Aset for speaker MHT of the ATR speech database [9]. Speaker MHT is a professional NHK narrator and has very clear, consistent speech. Clearly articulated speech was selected with the goal of emphasising poor concatenation of units so that the perceptual measure would be reliable. The Aset data from speaker MHT is also interesting because it is used by the current implementations of ATR ν-Talk and SUBPHONET and is one of the voices available for CHATR.

Isolated word pairs were selected from the database with one or more of their final phonemes being the same, so that the end of one word could be substituted into the other word leaving the phonemic string unchanged. With this method of construction, there is only a single concatenation point in each speech token. For a given word pair, the *common tail* is the maximum set of phones at the end of the words which is the same. For example, for /am*aeru*/ and /muk*aeru*/ the common tail is the last four phones /*aeru*/, as shown in Figure 1. The *splicing phoneme* is the first phoneme in the common tail (/*a*/ in this example). The *left context phones* are the phonemes which immediately precede the common tail (/*m*/ and /*k*/) and are guaranteed to be different. The *right context phones* immediately follow the splicing phoneme and are guaranteed to be the same for both tokens.

The set of 49 isolated word pairs shown in Table 1 was used in the experiment. Seven pairs were selected for seven splicing phonemes. The seven splicing phonemes were the five short vowels of standard Japanese, /*a,i,u,e,o*/ and two glides (/*w,y*/). These seven splicing phonemes were chosen because it has been found that sonorant phonemes are more difficult to concatenate smoothly than other phonemes. The criteria used in selecting the 49 isolated word pairs are described below in Section 2.1.1.

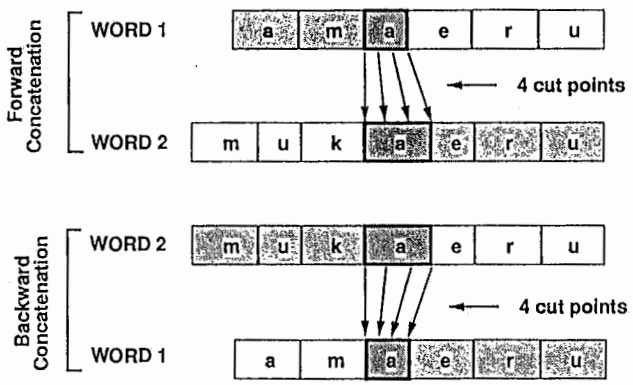

Figure 1: Determining the Token Cut-Points

Table 1: Word pairs used to produce the 490 speech tokens

Splicing Phone	Word 1	Word 2	Common Tail	Left Context
a	iwaba	cha	a	b, ch
a	maNga	fuda	a	g, d
a	shiwa	heiya	a	w, y
a	saka	haka	aka	s, h
a	saeru	atsuraeru	aeru	s, r
a	kawa	nawa	awa	k, n
a	amaeru	mukaeru	aeru	m, k
i	oite	tsuite	ite	o, u
i	taimeN	ichimeN	imeN	a, ch
i	jiei	suiei	iei	j, u
i	mimi	kimi	imi	m, k
i	nokogiri	deiri	iri	g, e
i	ichiou	taiou	iou	ch, a
i	iNki	niNki	iNki	-, n
u	narau	naibu	u	a, b
u	nurui	nikui	ui	r, k
u	mujuN	isshuN	uN	j, ssh
u	sue	yukue	ue	s, k
u	fuaN	zuaN	uaN	f, z
u	numa	uma	uma	n, -
u	ume	yume	ume	-, y
e	ue	me	e	u, m
e	hae	fue	e	a, u
e	seNi	keNi	eNi	s, k
e	hagemu	hohoemu	emu	g, o
e	chouetsu	eNzetsu	etsu	ou, z
e	geNiN	teNiN	eNiN	g, t
e	keNkyo	meNkyo	eNkyo	k, m
o	yo	go	o	y, g
o	fubo	ito	o	b, t
o	shio	mizo	o	i, z
o	kudoi	aoi	oi	d, a
o	hahaoya	chichioya	oya	a, i
o	kome	yome	ome	k, y
o	kotoni	omoni	oni	t, m
w	iwa	awa	wa	i, a
w	kawa	chouwa	wa	a, ou
w	majiwaru	yowaru	waru	i, o
w	kawari	owari	wari	a, o
w	kotowaza	shiwaza	waza	o, i
w	kuwawaru	majiwaru	waru	a, i
w	kouwa	kaiwa	wa	ou, i
y	sakuya	saya	ya	u, a
y	eiyuu	kiNyuu	yuu	ei, N
y	kayaku	keNyaku	yaku	a, N
y	nayamu	kuyamu	yamu	a, u
y	toNya	tokoya	ya	N, o
y	fuyu	yu	yu	u, -
y	nagoyaka	odayaka	yaka	o, a

For each isolated word pair, 10 speech tokens were produced, giving a total of 490 tokens. Two tokens were the natural recordings of the words (i.e. no concatenation involved). The natural recordings were included to determine the consistency of the human subjects in evaluating the quality of concatenation. The remaining tokens were produced by concatenating the common tail of word1 onto word2, *forward concatenation*, and the common tail of word2 onto word1, *backward concatenation*, as shown in Figure 1. In both these instances, four splice points were determined from the hand-produced phonemic labels provided with the database. The locations of the four concatenation points were:

- The start of the splicing phonemes,
- One third into the splicing phonemes,
- Two thirds into the splicing phonemes,
- The end of the splicing phonemes.

The concatenation of the tokens was performed by a splice close to the points specified above. The concatenation point could be shifted $\pm 3msec$ from the locations specified to avoid introducing discontinuities into the concatenated waveform. The exact concatenation point was selected to minimise the summed square of the difference in the waveforms over a seven sample window centered at the cut point.

2.1.1 Selection Criteria

The following criteria were used in the selection of the 49 isolated word pairs from amongst the hundreds of thousands of possible pairs:

- There should be some variability in the perceptual quality of the eight concatenated tokens for each word pair. This ensures some variability on which the signal processing estimate of the concatenation quality can be trained. Naturally, there was variability in concatenation quality for different word pairs. Thus, the token set had both within-word-pair variability and between-word-pair variability in concatenation quality.
- The intonation and durations of the common tail should be similar so that listener judgements are based primarily on the concatenation. This prevents tokens with good concatenation from being penalised because of an inappropriate f_0 contour. For example, take the first word as *sasayaka* which has accent type 2 when read in isolation (the morae are LHLL) and the second word as *kotosara* which has accent type 0 in isolation (the morae are LHHH). If we concatenate the common tail (/*a*/) from *kotosara* and use it to replace the final phoneme in *sasayaka* we get morae with a LHLH accent which is not acceptable as a non-marked production of an isolated word.
- There should be diversity in the left context phonemes. As Table 1 shows, almost all the pairs of left context phonemes are different[1].

[1]Note that the right context is controlled by the design of tokens from isolated words with a common tail.

2.2 Perceptual Tests

Nine native Japanese subjects evaluated the quality of the isolated word speech tokens described in the previous section. All subjects heard all 490 tokens one time each. The subjects were divided into 3 groups of 3 subjects. Each group listened to the tokens in different random orders in which no two tokens from the same word pair occurred in succession.

The tokens were recorded in advance onto a DAT tape with a Townshend Computer Tools DAT-Link resampling the speech data from 12kHz to 48kHz. Four second pauses were inserted between successive tokens. The 490 tokens were divided into 4 sessions, each taking approximately 10 minutes. There were 3 minute breaks between the sessions for the subjects to relax. The total experiment took around 50 minutes for each group. The subjects, seated in a low-noise room, listened to the tokens through STAX headphones (high-quality) with the volume adjusted individually for comfort at the start of the experiment.

Each subject had an evaluation sheet. The sheet listed the tokens in *hiragana* in order of presentation. The score sheet provided numbers from 0 to 10 which subjects circled to indicate the token quality from low quality (低音質) to high quality (高音質). The first 50 tokens were used for practice and have been ignored in all subsequent analysis. In a very small number of instances (2 out of 3960 results) a subject gave no response.

3 Initial Evaluation

An initial statistical evalation of the perceptual results was carried out to determine the level of agreement between the subjects, whether there was any effect for order of presentation, and other related statistics.

3.1 Inter-Subject Agreement

The perceptual scores for most of the subjects showed considerable variation. The standard deviations of their scores were in the range from 1.6 to 3.6 (as a reference, a uniform random distribution between 0 and 10 has a standard deviation around 2.8). This result indicates that the subjects did find substantial variation in the quality of the tokens.

The correlations between the perceptual scores for the different subjects were calculated. The correlations for pairs of subjects varied between 0.41 and 0.73 with a mean of 0.60 and median of 0.61 indicating moderate inter-subject agreement.

The mean of the subject responses for each token was calculated as a more robust estimate of the perceptual quality of the concatenations. The correlations between the scores of the subjects and the mean scores were considerably higher: the correlations varied from 0.67 to 0.89 with a mean of 0.80 and median of 0.82. This suggests that there is some variation in the individual scores (which is to be expected on a difficult perceptual task) but that the mean across all subjects is a reasonably reliable estimate of the perceptual quality of joins.

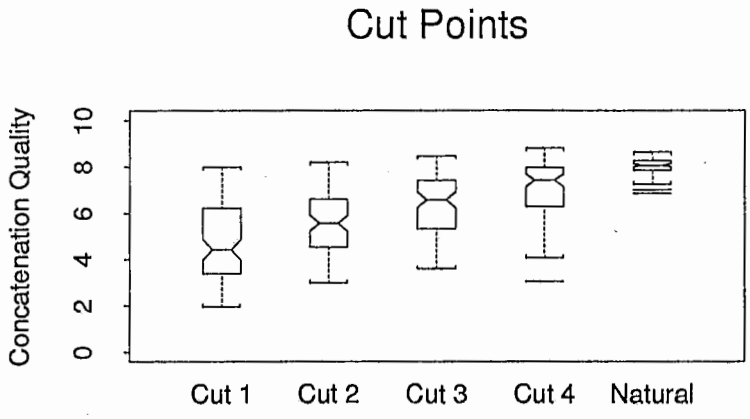

Figure 2: Comparison of Quality of Joins at Different Cut Points

Observation of individual scores suggested that some subjects had even distributions of scores across most of the range from 0 to 10, but that others had skewed distributions (for example, many low scores or high scores but relatively few in the middle). The following *annealing* technique was applied to the individual scores to compensate for this effect and produce a more robust combined estimate. For each subject, each response level (i.e. 0 to 10) was replaced by the mean of the cross-subject mean for the tokens with that response level. For example, if a subject gave a perceptual score of 4 for 14 tokens, we calcualte the mean response for all subjects for those tokens (say 4.3) and use that mean in ongoing calculations. These modified subject scores were then used to calculate a new cross-subject mean, which will be refered to as the *annealled mean.* This process was repeated until stable estimates were obtained (three iterations were sufficient). The annealled mean remained highly correlated ($r = 0.99$) with the cross-subject mean based on the raw scores but was more robust and was more readily predicted from the signal processing measures.

3.2 Effect of Cut-Point

As Section 2.1 described, four cut points were used to concatenate each word pair, and concatenated tokens were produced by forward and backward concatenation. Figure 2 shows a boxplot of the distribution of the concatenation quality from low to high (the annealled mean) for the five types of token; four concatenation points and the natural utterances.

Not surprisingly, the natural tokens were consistently scored as high quality (mean quality of 7.9 and standard deviation of 1.7). Also as expected, the average concatenation quality improved gradually as the cut point shifted from the start of the splicing phoneme to the end (from cut point 1 through to 4). This expectation was based on the fact that the left context phonemes of the two splicing phonemes are always different, while the right context phonemes are always the same. Therefore, there should be less coarticulatory difference in the later concatenation points. This result supports the general notion that units with

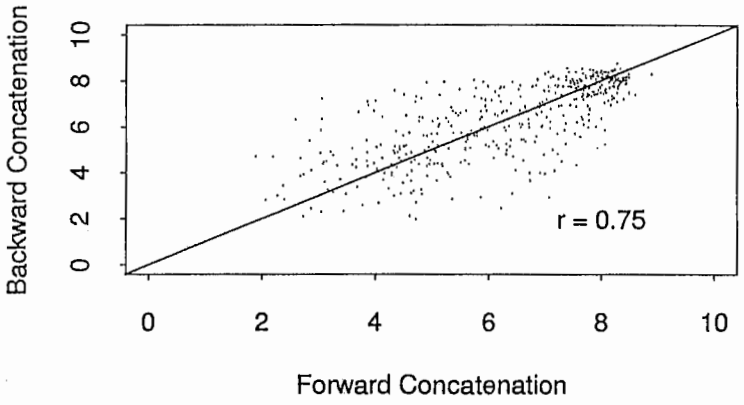

Figure 3: Comparison of Units with Reversed Concatenation

the same phonetic context are most likely to produce reasonable concatenation. However, because of the experimental construction we cannot conclude that concatenating at the end of a phoneme is better than concatenating in the middle or at the start of a phoneme. Moreover, it is not possible to determine how similarities or differences in phonetic context will affect concatenation (e.g., whether the effects of different voiced plosives on a following vowel are similar).

Figure 3 shows a plot of the annealled mean of the forward concatenation of a word pair versus the annealled mean for the backward concatenation of the same word pair at the same point. The correlation is 0.75. The reasonably high correlation suggests that at most concatenation points the direction of concatenation is not important. However, for some word pairs, there are substantial differences which suggests that there are dynamic effects or context effects which must be taken into account in predicting the perceptual quality. There is additional evidence for such effects from qualitative evaluation of matched forward and backward concatenation tokens. In some instances, though not many, the perception of the forward and backward concatenations could be substantially different though not necessarily better or worse. For example, the forward concatenation could be perceived in the high frequency discontinuity while the backward concatenation may be perceived at lower frequencies.

Perceptual masking is one possible explanation for this effect. However, the results presented in Section 4.2 show that dynamic representations of the spectrum (e.g. dynamic cepstra [10, 11] which is designed to capture perceptual masking) do not perform any better in predicting the perceptual quality of tokens in comparison to non-dynamic representations such as MFCC and power. This effect requires further study, possibly involving a much larger perceptual experiment.

4 Signal Processing

4.1 Non-Dynamic Signal Processing Parameters

A range of signal processing measures and combinations of these measures were used to predict the annealled mean perceptual scores for the tokens. The following are brief descriptions of the measures. Section 4.1.2 presents results from the analysis of these measures.

Mel-Frequency Cepstral Coefficients (MFCC):

The HCode program of Entropic's Hidden-Markov Model Toolkit (HTK) [8] was used to generate MFCC vectors [7] for each word used in the experiment. The following configuration was used:

- [-m] Produce MFCC vectors.
- [-p 24 -n 24] 24 x 24th order coefficients (24 MFCC parameters calculated from 24 Mel-frequency bands)[2].
- [-w 21.3] 21.3$msec$ window length (256 samples at 12kHz).
- [-f 5.0] 5.0$msec$ frame advance (200 frames per second).
- [-h] Hamming window.
- [-k 0.97] Pre-emphasis filtering.

The Euclidean distance between MFCC vectors nearest the cut-points[3] was used to predict concatenative quality. In the statistical evaluation, the number of MFCC parameters used in the calculation of the Euclidean distance was varied systematically between 1 and 24 (keeping the MFCC order fixed) as there was no *a priori* reason to assume that any specific number of parameters would be most effective or that increasing the number of parameters would improve the prediction. In the following sections, the notation $Dist^{n}_{MFCC}$ indicates the MFCC distance calculated as the Euclidean distance for the first n MFCC parameters as shown below, where $MFCC^{(1)}_{i}$ is the i_{th} MFCC parameter for the MFCC frame closest to the concatenation point in word1 (similar for $MFCC^{(2)}_{i}$ in word2).

$$Dist^{n}_{MFCC} = \sum_{i=1}^{n}(MFCC^{(1)}_{i} - MFCC^{(2)}_{i})^{2}$$

[2]20th order and 30th order MFCC parameters were also evaluated (in addition to 24th order). The results were very similar for all three settings.

[3]For all signal processing measures, the nearest vector will be selected. For features calculated with a 5$msec$ frame advance (MFCC, PCA of MFCC, delta MFCC, dynamic cepstrum) the maximum offset from the concatenation point is 2.5$msec$. For features calculated with a 10$msec$ frame advance (VQ of MFCC, improved cepstrum, power, f_0) the maximum offset from the concatenation point is 5$msec$.

Improved Cepstral Coefficients (ICep):

ATR ITL's `cepanaf` program was used to calculate improved cepstra for all words used in generating the concatenated tokens. This standard software produces 30 cepstral coefficients at $10msec$ intervals. Improved cepstrum uses an iterative smoothing process [6] which should provide a more accurate representation of the speech spectrum than conventional cepstral analysis [5]. As with MFCC, the number of improved cepstrum parameters was varied, in this case between 1 and 30. The cepstral distance provided by n improved cepstra will be represented by $Dist^n_{ICep}$ and calculated as:

$$Dist^n_{ICep} = \sum_{i=1}^{n} (ICep_i^{(1)} - ICep_i^{(2)})^2$$

Vector Quantisation of MFCC (VQ):

A major limitation with using spectral representations in on-line speech synthesis is that the storage and processing requirements for a large database are prohibitive. Vector quantisation of both MFCC and improved cepstrum vectors have been used in CHATR as a compact representation of the spectrum [3]. In the current work, the VQ parameters were produced by the Entropic ESPS `vqdes` program quantising the first seven MFCC parameters calculated with the `HCode` program (using the options given above). 256 VQ levels were used so that the spectral representation of each frame could be stored in a single byte. Using this method the spectral representation for an hour of speech (a large single speaker database) can be stored in less than half a megabyte. The distance between two VQ frames is determined from the distance between the prototype for each VQ level. The distance measure provided by the VQ MFCC vectors will be represented by $Dist_{VQ}$.

PCA Transformation of MFCC (PCA):

Principal Component Analysis (PCA) was evaluated as an alternative method to VQ for compressing the cepstral representation. PCA is a standard multivariate statistical technique which transforms a feature vector set to represent the variance in as few parameters as possible. PCA has been used previously in speech recognition to reduce a feature set size and has at the same time improved recognition accuracy [12]. In the current work PCA was applied to the first eight of the MFCC parameters (calculated as above) with the transformation weights determined from the complete Aset database of speaker MHT separately for each splicing phoneme. The distance measure calculated between n PCA MFCC parameters using the Euclidean distance will be represented by $Dist^n_{PCA}$ and was calculated as

$$Dist^n_{PCA} = \sum_{i=1}^{n} (PCA_i^{(1)} - PCA_i^{(2)})^2$$

Fundamental Frequency (f_0):

The squared difference of fundamental fequency was calculated from pitch tracks calculated by the `get_f0` program of ESPS with $10msec$ frame advance. This difference will be represented by $Dist_{fo}$.

Power:

Power was calculated for $20msec$ windows with $10msec$ frame advance using the Entropic ESPS `get_f0` program. The squared differences in the raw power and the log of power were calculated; these will be represented by $Dist_P$ and $Dist_{logP}$ respectively. Similarly, the absolute value of the differences in the raw power and the log of power were calculated; these will be represented by $AbsDist_P$ and $AbsDist_{logP}$ respectively.

4.1.1 Modifications to the Distance Calculation

Two modifications to the distance measures described above where evaluated to determine whether they would effect the prediction of concatenation quality.

Distance of Raw to Concatenated Units:

When concatenating the waveforms of two units, it is possible that the spectrum in the concatenated waveform at the join point might not be close to the spectrums of the waveforms being concatenated or to an interpolation between those two spectrums[4]. The possibility of such spurious spectral characteristics occurring has been postulated as one possible cause of poor concatenation quality in concatenative synthesis. If this were to be the case, then the distance between the spectrums of the original units and the concatenated waveform at the concatenation point should be a good predictor of the quality of concatenation. However, in all cases, this predictor was found to be substantially *worse* than distances based on the pre-concatenation representations. Thus, there seems to be little evidence for this proposition.

From a practical viewpoint, this result is very good for speech synthesis. It means that in the training and the operation of concatenative speech synthesis systems we do not need to concatenate and signal process units to obtain the best estimates of concatenation quality. Instead, all calculations can be based on spectral representations derived from a single pass of signal processing of a corpus.

Absolute Value of Differences:

In addition to using the Euclidean distance between the parameters from two words (sum of squares of the differences), the sum of the absolute differences was evaluated for each of the spectral measures described above. In some cases the distance measure using absolute values provided similar predictive accuracy to the Euclidean distance but in other cases, the performance was somewhat poorer. Thus, for consistency, the results presented here use the Euclidean distance. Only for the power and the log of power measures was the absolute difference a significantly better predictor than the squared difference and thus both measures are presented in the next section.

[4]This may result from an artefact of windowing the waveforms when concatenating, for example, from the square window used in the current work. In an overlap and add system, e.g. PSOLA, there is less chance of such artefacts occurring. Nevertheless, the results of the current work suggest that direct cutting (not overlap) can be effective if suitable points on the waveforms are identified.

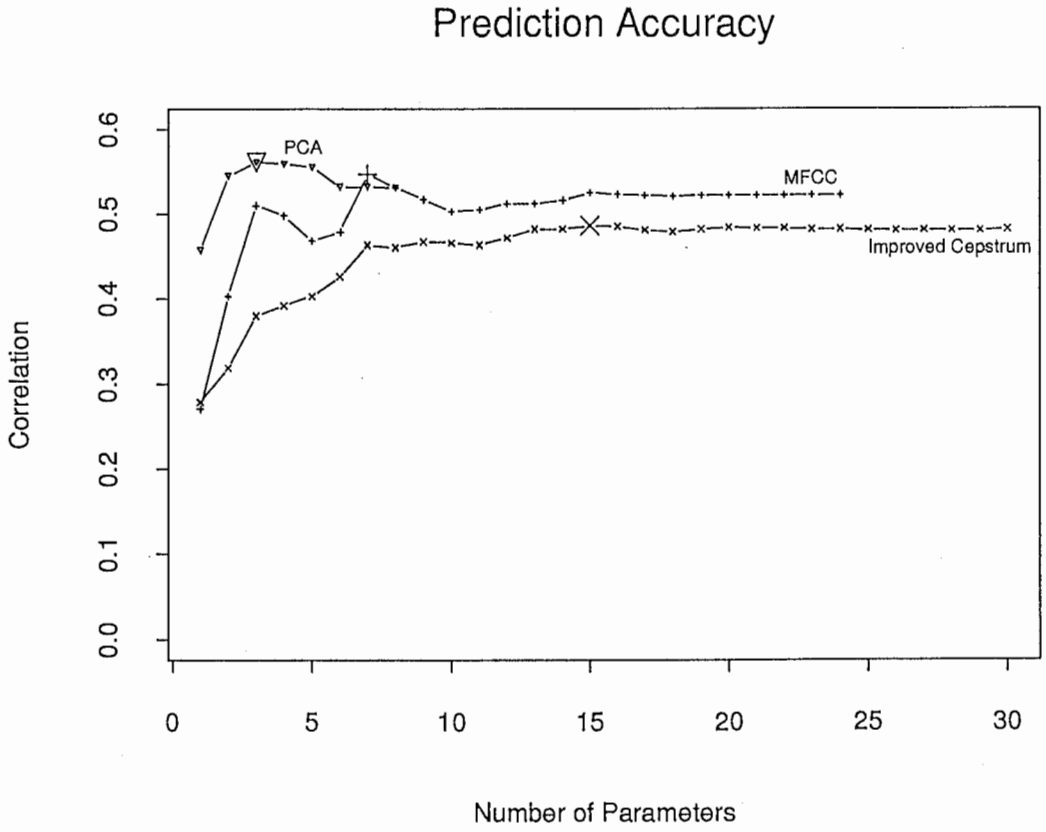

Figure 4: Prediction of the concatenation quality for MFCC, improved cepstrum and PCA MFCC for varying numbers of parameters

4.1.2 Predicting Perceptual Quality

The correlation between the series of distance measures described above and the annealled mean was used as an indication of effectiveness. Figure 4 shows a plot of this correlation for $Dist^n_{MFCC}$ for n from 1 to 24, $Dist^n_{ICep}$ for n from 1 to 30, and for $Dist^n_{PCA}$ for n from 1 to 8. The large marks indicate the best performance for the different spectral representations.

The best prediction is provided by the Euclidean distance between the first 3 PCA MFCC parameters which has a correlation of 0.561. The best MFCC prediction occurs with the Euclidean distance between the first 7 MFCC coefficients and has a correlation with the annealled mean of 0.547 (2.5% below the correlation for PCA). The prediction accuracy using 24 MFCC parameters is slightly lower at 0.522. The best prediction using the improved cepstrum is 0.486 (13% lower than for PCA) and is the Euclidean distance of the first 15 parameters. Using all 30 improved cepstrum parameters, as ATR ν-Talk does, there is a very slight drop in predictive accuracy to a correlation of 0.482.

The figure shows that MFCC parameters are consistently more effective than the improved cepstrum and that a small number of PCA parameters derived from MFCC are much better than small numbers of MFCC or improved cepstrum parameters. The result

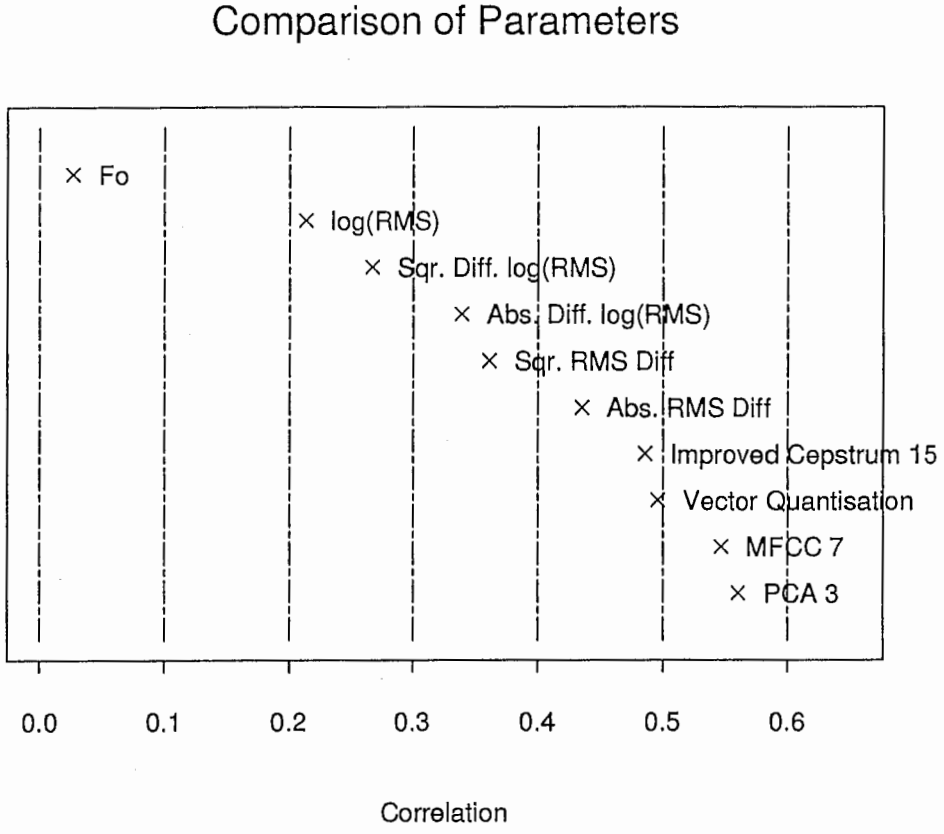

Figure 5: Comparison of Spectral Representations

that the prediction of concatenation quality can be improved by reducing the number of parameters is important as it indicates that storage and processing requirements for spectral representations can be reduced while still improving concatenation quality.

Figure 5 plots the predictive accuracy (as a correlation to the annealled mean) for the range of signal processing parameters. Table 2 presents the same results in table form. In general terms, the spectral representations (MFCC, PCA of MFCC and improved cepstrum) are more effective than the basic acoustic prosodic measures (f_0 and various representations of power). The best spectral representation for predicting perceptual quality uses 3 PCA parameters. Next best is 7 MFCC parameters. Vector quantisation of MFCC is next best and it marginally exceeds the predictive accuracy of 15 (or 30) improved cepstrum parameters.

The f_0 difference provides very poor prediction of the concatenation quality ($r = 0.027$). This is likely to be because the design of the experimental tokens controlled the intonation of the units. Without substantial differences at the cut points, it would be unlikely for f_0 to be a useful predictor. A different experimental configuration should be used to evaluate the contribution of f_0 to the perception of concatenation.

The five parameters derived from instantaneous power were reasonable predictors. In-

Table 2: Correlations of Spectral Representations

Feature	Correlation	Dimen.
f_0	0.027	1
log of rms energy	0.213	1
Squared difference of log of rms power	0.267	1
Absolute value of difference in log of rms power	0.339	1
Squared difference of rms power	0.361	1
Absolute value of difference in rms power	0.435	1
Euclidean distance of 30 improved cepstrum parameters	0.482	30
Euclidean distance of 15 improved cepstrum parameters	0.486	15
Vector quantisation distance	0.496	1 Byte
Euclidean distance of 7 MFCC parameters	0.547	7
Euclidean distance of first 3 PCA parameters	0.561	3

terestingly, the instantaneous power in the unit immediately prior to the cut-point was significantly correlated with the concatenation quality ($r = 0.213$). This suggests that louder units were slightly more likely to have poorer concatenation. A number of reasons have been considered for this result but none can be confidently supported by the current results. The difference in rms power is a better predictor than the difference in log of rms power: this provides further support for the result that poor concatenation in louder units is more perceptually salient. Also, the absolute value of the difference in the power measures is more effective than the squared difference. Thus, the absolute value of the difference in rms power (without log) is the best predictor amongst the power terms ($r = 0.435$).

4.1.3 Storage Requirements

For practical purposes, it is important to consider the storage space requirements for the different signal processing representations. In the current work all features except the vector quantisation were represented by single precision floating point numbers (4 bytes per value). The basic f_0 and power features use a single value per frame, and the spectral representations used the numbers of values shown in the final column of Table 2. The table allows us to make considered judgements of how best to trade off predictive accuracy with storage and processing requirements. Both the vector quantisation representation and the PCA of MFCC are effective predictors with minimal storage requirements.

A comparison of the VQ of MFCC and Euclidean distance for 30 improved cepstrum parameters (used in ATR ν-Talk and SUBPHONET) is particularly interesting. VQ of MFCC provides 120 times reduction in storage requirements with no significant change in predictive accuracy (in fact with a very slight improvement).

4.2 Dynamic Features

All six measures described above are measures of a speech waveform at a particular point in time. It is well known that the dynamic properties of the speech waveform are also very important to the perception of speech. For this reason, variants of the spectral parameters which captured dynamic (or delta) characteristics were evaluated to determine whether they can contribute to the prediction of the perceptual quality of unit concatenation. The results were consistent and surprising; none of the dynamic features improved the prediction of concatenation quality, and in fact, most degraded the predictive accuracy in comparison to their non-dynamic counterpart.

Delta-MFCC:

The Euclidean difference between delta-MFCC parameters for the two tokens at the cut-point was a poor predictor of concatenation quality. For Euclidean distances taken from 1 through 24 delta-MFCC parameters, the correlation to the annealled mean is less than 0.05.

Dynamic Cepstrum:

The dynamic cepstrum representation of speech [10, 11] attempts to capture the spectral masking effect of human audio perception. It can be viewed as a combination of standard cepstrum and delta cepstrum and it has broad similarities to RASTA processing [13]. In the current work, dynamic cepstrum parameters were derived from the MFCC parameters described earlier in this report. The controlling parameters of the dynamic cepstrum defined by Aikawa *et al.* [11] $(\alpha, \beta, q_0, \nu)$ were varied over a wide range and the resulting parameterisation was used to predict concatenation quality. In addition, both square and Gaussian windows were implemented. In almost all cases, the dynamic ceptrsum performed substantially worse than the MFCC parameters from which they were derived. Moreover, the few conditions in which little decrease occurred or slight improvement occurred (of less than half a percent) had settings which provided almost no dynamic component to the model (e.g. $\alpha = 0.02$). In short, dynamic cepstrum, like delta-MFCC appeared to be of little benefit for predicting concatenation quality.

4.3 Predicting by Combining Measures

All results presented so far in this report were for prediction of the concatenation quality by single parameters. Linear combinations of parameters were also evaluated to determine whether they could perform better than the individual parameters. The combined predictive model is of the form:

$$ConcatenationQuality \leftarrow \sum_i w_i * f_i + c$$

where f_i is a selection of the features evaluated in the previous sections, and w_i are the

Table 3: Prediction by Combining Spectral and Power Measures

Feature	Prediction Correlations			
	Original	+ $Dist_P$	+ $AbsDist_P$	+ Both
$Dist^{15}_{ICep}$	0.486	0.525	0.566	0.591
$Dist_{VQ}$	0.496	0.536	0.573	0.594
$Dist^{7}_{MFCC}$	0.547	0.565	0.598	0.630
$Dist^{3}_{PCA}$	0.561	0.582	0.616	0.648

weights for those features. The intercept term, c, is not relevant to the current work and no results will be reported for it.

For practical reasons, it is unlikely that multiple spectral representations would be used together in a concatenative speech synthesis system (because of storage and processing requirements). Thus, only combinations of the power terms and the four spectral representations were investigated.

Table 3 shows the prediction accuracy (as a correlation) for regression models trained with combinations of the four spectral representations and power terms. The table provides results for 16 regression models. For each of the four spectral representations, improved cepstrum, VQ of MFCC, MFCC, and PCA of MFCC four regression models were constructed:

Original: The spectral measure alone. This provides the same result as that given in Table 2.

+$Dist_P$: Combination of the spectral measure and the squared difference in power at the cut point.

+$AbsDist_P$: Combination of the spectral measure and the absolute value of the difference in power at the cut point.

Both (+$Dist_P$ + $AbsDist_P$): Combination of the spectral measure and the absolute value and square of the difference in power at the cut point.

The addition of $Dist_P$ improves the correlations by between 0.02 and 0.04 (from 3% to 8%). The addition of $AbsDist_P$ improves the correlations by between 0.05 and 0.08 (from 9% to 16%). Finally, the addition of both the power terms to the spectral representations increases the correlations by between 0.08 and 0.10 (from 15% to 22%). Clearly, the addition of the power terms can provide substantial (and significant) improvements in the predictive accuracy.

From a practical viewpoint, the addition of the two power terms requires only one additional value to be stored for each frame in a speech corpus. This increase in storage and

processing requirements should be offset by improvements in the speech quality. The trade-off between predictive accuracy and storage requirement discussed in Section 4.1.3 remains valid. The combination of power with either VQ of MFCC or PCA of MFCC are compact representations with reliable prediction of concatenation quality and are therefore well-suited to real-world speech synthesis systems.

The final issue to consider is the relative contribution of the spectral and power terms to the regression models; in other words, the values for the weights w_i. The intrepretation of the weights is difficult as they are dependent upon the way in which the spectral and power measures are calculated. Indicative measures can be obtained by z-score normalisation of the measures prior to training of the regression model.

For the model $Dist^7_{MFCC} + AbsDist_P$, the weights are 0.75 and 0.24 for the MFCC and power measures respectively (the coefficients are both positive). In other words, the MFCC term is about three times more important[5]. The pattern is similar for the other spectral representations.

The predictive accuracy is relatively insensitive to the weights used. For example, for a model $Dist^7_{MFCC} + AbsDist_P$ with weights 0.5 and 0.5 (instead of 0.75 and 0.24) provides predictive correlation of 0.589 which is only 1.5% below the maximum obtainable. Similarly, weights of 1.0 and 0.24 provide predictive correlation of 0.585 - 2.1% below the maximum.

5 Discussion

The execution of the perceptual experiment has permitted empirical evaluation of a range of signal processing measures as estimates of the quality of concatenation of units for speech synthesis. The result that MFCC out-performs the improved cepstrum is not surprising given the psychoacoustic basis of the Mel frequency scale and the consistent effectiveness of MFCC in speech recognition. The difference is only around 12%.

The effectiveness of the compressed representations of MFCC provided by VQ of MFCC and PCA of MFCC is particularly important. The one-byte representation (per frame) of VQ provides predictive accuracy slightly above that of 30 floating point improved cepstrum parameters despite the 120 times reduction in storage requirment. The PCA transform of MFCC produces 3 floating point parameters which have slightly better predictive accuracy than the 7 untransformed MFCC parameters with more than 50% reduction in data.

It would be trivial to quantise any of the three spectral measures which were represented by single-precision floats in the current work (improved cepstrum, MFCC and PCA of MFCC). This could be achieved by scaling into the range 0-255 and quantising to the nearest integer. With this quantisation a single byte could replace the floating point numbers - a four times data reduction. It is unlikely that this procedure would impact upon the predictive accuracy and should be seriously considered for the implementation of computationally

[5]However, as the MFCC distance was derived as the sum of seven MFCC parameters, the power term is about twice as important as each of the individual MFCC parameters.

effecient synthesis systems.

An issue not addressed by the current work is the extent to which time resolution of the signal processing measures affects the prediction of concatenation quality. For example, is there any difference between MFCC parameters with $5msec$ and $10msec$ frame advance? The expectation (without any empirical study or support) is that better time resolution should improve the predictive accuracy up to a point. It is likely that there would be a trade-off between the increased storage requirments with finer time resolution and the expected improvement in accuracy.

The role of power in a concatenative speech synthesis system is special because power is a signal characteristic which can be trivially modified in the processing of a waveform. For example, concatenation by PSOLA [14] permits support for frame-by-frame manipulation of power. In the current experiment, no such manipulation of power was used; the isolated word segments were concatenated without modification and without overlap (note that appropriate points on the waveforms were selected to minimise discontinuity). The results regarding prediction of concatenation quality from the spectral measures are unaffected. What must be considered is whether the combination of spectral measures and power, which improved the prediction of concatenation quality (Section 4.3), can improve the selection of units in a concatenative speech synthesis system. There are theoretical arguements both for and against the inclusion of power - this is an issue which must be resolved by investigation on an operational speech synthesiser.

The current work is based on the perception of joins in seven different phonemes, the five short vowels of Japanese /a, i, u, e, o/ and two glides /y, w/. We must consider the issue of whether the current results are applicable to other phonemes. It is very likely that the results will translate well to long vowels as they are spectrally similar to the short vowels. Similarly, the spectral distances are likely to be reliable for nasals. By comparison, we have no reason to believe that the results will be applicable to other consonants of Japanese (voiced and unvoiced fricatives, voiced and unvoiced plosives, affricates etc). In the absence of experimental evidence there is no choice but to use the existing spectral measures which this experiment has shown are reasonable.

Another issue to consider is whether the results presented here are applicable to other languages, in particular English which is the second language currently synthesised by CHATR. In pilot experiments to prepare the speech tokens used in the current work, the first author (native English speaker and a poor speaker of Japanese) and a native Japanese speaker both evaluated several hundred Japanese speech tokens using a method very similar to that described in Section 2.2. There was very good agreement between the perceptual scores (correlation of 0.80). This suggests that the perception of concatenation quality may be reasonably invariant across languages, but no strong conclusion can be drawn. Once again, in the absence of experimental evidence it is sensible to use the same estimate of concatenation quality for both English and Japanese.

5.1 Application to CHATR

As the results of the current work were obtained, several modifications were made to the CHATR synthesiser and consistent improvements were obtained. Firstly, the improved cepstrum was replaced by MFCC parameters in the training of CHATR, and in the calculation of the concatenation cost used to select units. Subjective evaluation by the second author indicated that this change provided consistent, but not dramatic, improvement in the quality of synthesis. Secondly, the number of MFCC parameters was reduced from 12 to 7. This did not appear to affect synthesis quality but reduced storage requirements substantially. Thirdly, a combination of power and MFCC was used to calculate concatenation costs. This provided quite significant improvement in speech quality.

The successful application of the results of the current research to CHATR indicates that the theoretical findings from the perception experiment do indeed translate to practical improvements. Moreover, the results have also greatly assisted ongoing work by the authors on the automatic training of the CHATR system. From a methodological viewpoint, we have found that there is merit to theoretical investigation of many of the "intuitions" that guide speech synthesis system development.

6 Conclusion

The perception experiment presented in this report has provided subjective judgements of a substantial number of speech tokens. This has provided the resource to evaluate a number of potential signal processing measures for predicting the concatenation quality. The experiment has confirmed that the cepstrum distance is reasonably effective in predicting concatenation quality and that Mel-frequency cepstral coefficients are better than improved cepstra. From a more practical viewpoint, the results show that compression of MFCC parameters using vector quantisation or principal component analysis can substantially reduce the storage and processing requirements for a speech synthesis system without greatly affecting the accuracy in the prediction of concatenation quality.

Improved predictions were obtained using linear regression to combine power difference measures with the spectral distances. The best predictive accuracy was obtained by linear weighting the Euclidean distance between the first 3 PCA MFCC parameters with the absolute value of the difference in power and the squared difference in power. The prediction by this combination had a correlation with the perceptual score of 0.65 which compares favourably with a correlation of 0.48 obtained with 30 improved cepstrum parameters (the method previously used in CHATR and currently used in ATR ν-Talk and SUBPHONET). The incorporation of this result into CHATR has improved its speech quality.

With the best predictive accuracy being 0.65, there is clearly potential for improvement. Unfortunately, there are no obvious paths for improvement. However, if a new idea is developed then the speech tokens and perceptual scores produced by the current work remain available for researchers. The availability of such data should facilitate much more rapid

evaluation of new estimates of concatenation quality.

6.1 Future Work

- The PCA of MFCC distance measure will be implemented in CHATR.
- Perceptual evaluation of the concatenation of a complete range of phonemes needs to be performed to determine a complete picture of the applicability of the MFCC distance and other measures.
- It is possible that listeners may be more sensitive to concatenation in particular areas of speech (e.g. stressed vowels in English). A larger experiment may consider this issue.
- The four spectral measures used in the current work were calculated using a Euclidean distance. The Mahalanobis distance might be considered as an alternative.
- Perceptual evaluation of a large number of concatenation points in the *actual* output of a speech synthesiser may be considered as it may indicate biases in the unit selection algorithms.

Acknowledgements

Many people assisted with the execution of the work presented here. The authors are particularly grateful to Toshio Hirai for his help with translations and for verifying the intonation of the experimental tokens, and to Dr Hiroaki Kato for assistance with the design of the perceptual experiment. The discussions with Drs. Norio Higuchi, Nick Campbell, Yoshinori Sagisaka and Kiyoaki Aikawa helped in formulating the experiment. Our thanks go to the members of Department 2 of ATR ITL for participating as subjects in the perceptual experiment. The authors are grateful to Dr. Yasuhiro Yamazaki for his ongoing support.

References

[1] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura. ATR v-talk speech synthesis system. In *Proc. 1992 Intl. Conf. on Spoken Language Processing*, pages 483–486, Banff, Canada, 1992.

[2] N. Iwahashi, N. Kaiki, and Y. Sagisaka. Concatenative speech synthesis by minimum distortion criteria. In *ICASSP '92*, pages II–65–68, 1992.

[3] A. Black and N. Campbell. Optimising selection of units from speech databases for concatenative synthesis. In *EUROSPEECH '95*, pages 581–584, Madrid, Spain, 1995.

[4] N. Campbell and A. 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 Verlag, 1995.

[5] D. Childers, D. P. Skinner, and R. C. Kemerait. The Cepstrum: A guide to processing. *IEEE Proceedings*, 65:1428–1443, 1977.

[6] S Imai and Y Abe. Spectral envelope extraction by the improved cepstral method. *Trans. ICICE*, J63-A(12):217–223, 1979.

[7] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980.

[8] S.J. Young, P.C. Woodland, and W.J. Byrne. Htk: Hidden markov model toolkit v1.5 user manual. Technical report, Entropic Research Laboratories Inc., 1993.

[9] K. Takeda, Y. Sagisaka, S. Katagiri, M. Abe, and H. Kuwabara. Speech database user's manual. Technical report, ATR Interpreting Telecommunications Research Laboratories: TR-I-0166, 1988.

[10] K. Aikawa and R.A. Yamada. Comparative study of spectral representations in measuring the English /r/-/l/ acoustic-perceptual dissimilarity. In *Proc. Intl. Conf. on Spoken Language Processing*, pages 2039–2042, Yokohama, Japan, 1994.

[11] K. Aikawa, H. Singer, H. Kawahara, and Y Tohkura. A dynamic cepstrum incorporating time-frequency masking and its application to continuous speech recognition. In *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing*, pages II–668–671, 1994.

[12] A. J. Hunt and R. Favero. Using principal component analysis with wavelets in speech recognition. In *Proc. 5th Aust. Intl. Conf. on Speech Science and Technology*, pages 296–301, Perth, Australia, 1994.

[13] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Trans. on Speech and Audio Processing*, 2(4):578–589, 1994.

[14] E. Moulines and Charpentier F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5/6):453–467, 1990.