APPLICATION OF THE DYPSA ALGORITHM TO SEGMENTED TIME SCALE MODIFICATION OF SPEECH

Mark R. P. Thomas, Jon Gudnason and Patrick A. Naylor

Imperial College London
Exhibition Road, London SW7 2AZ, UK
E-mail: {mrt102, jg, p.naylor}@imperial.ac.uk

ABSTRACT

This paper presents a method for speech time scale modification. Voiced speech is pseudo-periodic, allowing time scale modification by the repetition or removal of cycles as necessary. However, in the case of unvoiced speech and at the boundaries of voiced speech, no such periodicity exists so the speech should not be modified. To address this issue, the proposed approach is novel in its use of the DYPSA algorithm to derive speech periodicity from glottal closure instants (GCIs), followed by a Gaussian Mixture model-based voiced/unvoiced/silence (VUS) classifier. A listening test based on ITU-T P800 has been conducted and has shown that, by employing VUS detection, the average mean opinion score of the perceptual quality of processed speech exceeds that of a method without VUS detection by 0.61 over a range of modification factors. Results are presented as a function of modification factor for normal and fast original talking rate. Reliable time scale modification of high audio quality enables many applications, such as time scale compression for fast scanning of recorded voicemail messages, slowing talking rate for improved intelligibility in forensics and lip synchronization in motion video.

1. INTRODUCTION

Speech time scale modification is a process which alters the length of a segment of speech without significantly affecting its pitch or formant structure. It has many uses, including time scale compression for fast scanning of recorded voicemail messages [1] and time scale expansion for improving the intelligibility of fast or degraded speech in forensic applications. A combination of compression and expansion may also find uses in the synchronization of audio to lip movements in motion video. Real-world applications of time scale modification have, however, been limited due to the presence of unwanted artefacts in existing approaches. This paper presents a new approach which reduces many common artefacts and provides fast and perceptually superior results.

During voiced speech, the pseudo-periodicity of the waveform naturally lends itself to time scale modification as complete larynx cycles may be removed or repeated depending upon whether a compression or expansion of signal duration is desired. Providing that the periods are accurately known and cycles are concatenated in such a way that pitch periods are faithfully reproduced, good time scale modification can be achieved. However, during periods of unvoiced speech, voiced fricatives, plosives or boundaries of voiced speech, no such periodicity exists, though most algorithms still apply uniform time scale modification to the entire speech signal. These segments will be referred to hereon as *unvoiced and transition* (UT) segments. The resulting artefacts in UT segments, caused by algorithms such as the following, diminish the quality of the processed speech.

Existing approaches for concatenating periods of voiced speech for time scale/pitch modification include the PSOLA method [2] and specifically time-domain PSOLA (TD-PSOLA) which performs well provided a) pitch periods are accurately known and b) high quality time scale (but not pitch) modification is required. Other approaches include sinusoidal-based [3], LP residual-based (LP-PSOLA) [4, 5], waveform similarity-based (WSOLA) [6] and phase vocoders [7], which address the cases when one or more of these constraints are unfeasible, at the cost of added complexity.

More recent approaches address the issue of UT segments [8, 9, 10, 11, 12]. In the literature, effort has been made to apply different levels of duration modification for different segments with positive results but little work has been done to optimize these parameters. The studies generally conclude that the most perceptually significant artefacts are those arising from the repetition of UT segments, for which the fast and accurate detection is key in the proposed algorithm. The approach also differs in the use of the DYPSA algorithm [13] to quickly and reliably find glottal closure instants (GCIs) to use as pitch markers.

The strategy is to address the problem of modification during segments of little or no periodicity by employing a voiced/unvoiced/silence (VUS) detector, from which UT segments are derived. It assumes that the duration of most UT segments is independent of speech rate [14, 10] and does not apply modification to them. During voiced segments, DYPSA provides GCIs which are used as pitch markers. During silence, the algorithm places pseudo-pitch markers every 10 ms. Cycles are then concatenated using PSOLA, ensuring that pitch periods are faithfully reproduced using the approach in [5]. The result is a practical, fast and reliable method for time scale modification that is novel in a) the use of DYPSA to find pitch markers and b) the use of a Gaussian Mixture-based classifier to find UTs. Subjective testing has shown that the proposed method gives significantly greater mean opinion scores than an equivalent method which performs uniform processing on the entire speech signal.

This paper is organised as follows: Section 2 formulates the problem with a set of examples. Sections 3 and 4 describe the DYPSA algorithm and the VUS detector respectively. Results and discussion of subjective tests are presented in Section 5 followed by conclusions in Section 6.

2. MOTIVATION FOR THE PROPOSED APPROACH

Compression or expansion of speech time scale involves removing or repeating cycles as required, as shown in Fig. 1. Pitch markers must be pitch-synchronous, but by identifying GCIs, it is guaranteed that in addition to being pitch-synchronous, crossfades take place where there is low speech energy. Such an approach can give good time scale modification during periods of voiced speech providing the GCIs are accurate. The DYPSA algorithm ensures GCIs are accurately estimated, eliminating the 'phasiness' property which often accompanies poor GCI estimation in time scale modification [7]. This is a highly important case for forensic applications where speech is often slowed down in order to ease the task of transcription.

However, Fig. 2 shows that by using this approach with pseudo pitch periods placed every 10 ms during unvoiced sounds, the waveform can acquire periodic components it did not originally possess when the time scale is stretched. This gives a very unnatural-sounding result which diminishes the overall quality of the processed speech. In the case of a fast-spoken sentence which is to be slowed down, there is a greater ratio of the duration of unvoiced to the duration of voiced speech, as the duration of many

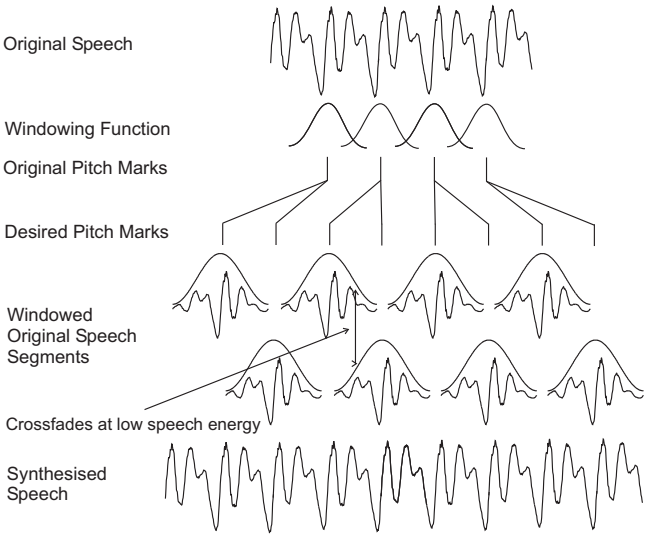

Figure 1: Concatenation of pitch periods. Periodicity is identified (in this case at the instants of glottal closure) and individual periods are multiplied with a Hamming window. Periods are repeated or removed as necessary and then aligned and normalised to form a new synthesised signal with modified time scale. The use of GCIs as pitch markers ensures that crossfades take place during regions of low speech energy.

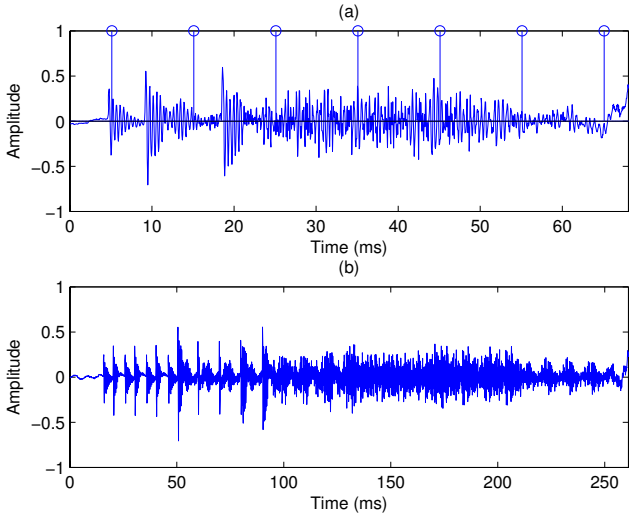

Figure 2: The effect of treating unvoiced sounds as periodic in time expansion. Speech signal (a) is an utterance of the phoneme/tʃ/ (containing both impulsive events and turbulent noise), with pseudo pitch markers placed every 10 ms. A time expansion of four times is shown in (b) which contains many additional harmonic components.

unvoiced sounds has been found to be largely independent of talking rate [14, 10]. The aforementioned artefacts will therefore be worse in the case of time scale expansion on fast-spoken speech.

When compressing the time scale of a speech signal, a uniform approach can cause short (but important) sections of speech to be removed altogether and significantly impair intelligibility. Fig. 3 gives an example where a plosive is lost.

These problems may be addressed if UT segments are isolated and left unchanged, allowing only voiced and silent periods to be modified. The assumption that many unvoiced sounds are either weakly proportional to talking rate or are entirely independent, is mentioned in [14, 10] and backed up by subjective testing in Section 5.

3. THE DYPSA ALGORITHM

The main features of the Dynamic Programming Phase Slope Algorithm (DYPSA) are now reviewed. It consists of three main components: the phase slope function, phase slope projection, and dynamic programming. These components are defined as follows.

Phase-slope function [15] – defined as the average slope of the unwrapped phase spectrum of the short time Fourier transform of the prediction residual. GCI candidates are selected based on the positive-going zero crossings of the phase-slope function.

Phase-slope projection – introduced to generate GCI candidates when a local minimum is followed by a local maximum without crossing a zero. The midpoint between these is identified and projected onto the time axis with unit slope. In this way, GCIs whose positive going slope does not cross the zero point (those missed by the phase-slope function) are identified.

Dynamic Programming – uses known characteristics of voiced speech and forms a cost function to select a subset of the GCI candidates which are most likely to correspond to the true ones. The subset of candidates is selected according to the minimisation problem defined as

$$\min_{\Omega} \sum_{r=1}^{|\Omega|} \lambda^T \mathbf{c}_{\Omega}(r), \qquad (1)$$

where Ω is a subset with GCIs of size $|\Omega|$ selected from all GCI candidates, $\lambda = [\lambda_A\ \lambda_P\ \lambda_J\ \lambda_F\ \lambda_S]^T = [0.8\ 0.5\ 0.4\ 0.3\ 0.1]^T$ is a vector of weighting factors with the values taken here as in [16] and $\mathbf{c}(r) = [c_A(r)\ c_P(r)\ c_J(r)\ c_F(r)\ c_S(r)]^T$ is a vector of cost elements evaluated at the rth GCI of the subset. The cost vector elements are:

- *Speech waveform similarity*, $c_A(r)$, between neighbouring candidates, where candidates not correlated with the previous candidate are penalised.
- *Pitch deviation*, $c_P(r)$, between the current and the previous two candidates, where candidates with large deviation are penalised.
- *Projected candidate cost*, $c_J(r)$, for the candidates from the phase-slope projection, which often arise from erroneous peaks.
- *Normalised energy*, $c_F(r)$, which penalises candidates that do not correspond to high energy in the speech signal.
- *Ideal phase-slope function deviation*, $c_S(r)$, where candidates arising from zero-crossings with gradients close to unity are favoured.

It can be seen from the dynamic programming criteria that DYPSA is robust to spurious peaks in the prediction residual, providing a GCI detection rate of $\sim 96\%$ during voiced speech. The use of PSOLA in conjunction with DYPSA – which relies heavily on waveform similarity – is loosely related to the WSOLA technique [6], with the notable exception that DYPSA operates on a set of candidates derived from the LPC residual with the Group Delay method.

4. SPEECH SEGMENTATION

The purpose of the VUS detector is to segment a speech signal into three classes: voiced, unvoiced and silent. During voiced segments, the DYPSA algorithm [13] is used for GCI detection. During silence, pseudo pitch markers are placed every 10 ms. Time scale modification may then be applied to the marked voiced and silent segments; UT regions are left unprocessed. VUS detection and the derivation of UT segments are described in the remainder of this section.

4.1 VUS Probabilities

VUS detection is based on feature vectors derived from 20 ms frames of speech. Each class is modelled by a multivariate full covariance Gaussian distribution, whose parameters are derived from

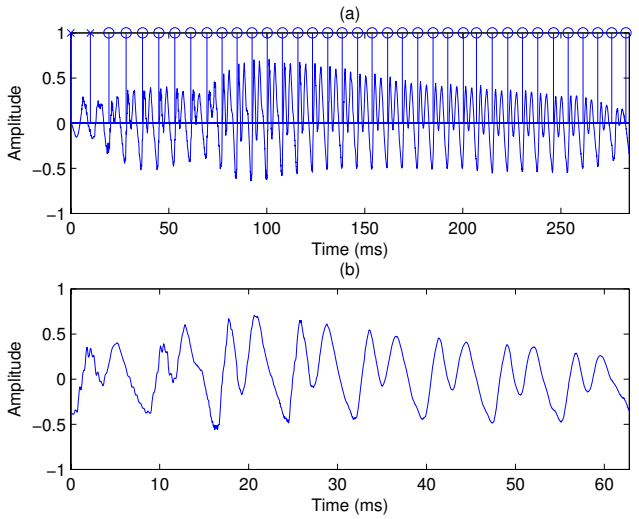

Figure 3: The effect of treating unvoiced sounds as periodic in time compression. Speech signal (a) is an utterance of the word /blu/ (blue), with pseudo pitch periods marked '×' and GCIs marked '○'. A time compression of four times is shown in (b) which sounds closer to /lu/.

labelled training data. Once trained, the VUS detector is applied to a set of test data for the forthcoming subjective tests.

The five feature vector coefficients are the well-known parameters defined in [17]. *Zero crossing rate* indicates how the energy of the speech signal is concentrated in frequency. Voiced segments of speech have a low zero crossing rate. This measure varies significantly for silent periods because of its dependence on ambient noise. *Normalized Energy* is greatest during voiced segments of speech. *Normalized autocorrelation coefficient at 0.1s delay* is a strong indicator of the spectral whiteness of unvoiced speech whose value is close to zero. The periodicity of voiced speech gives a value close to one. *Mean spectral slope*, estimated by the first covariance LPC coefficient, is much steeper for voiced speech due to the nature of the glottal volume velocity exciting the vocal tract. *Energy in LP residual* is a good indicator of the strength of formants present in voiced segments of the speech signal. The five coefficients are then augmented with their delta and delta-delta coefficients to produce a 15 dimensional feature vector $\mathbf{x}_i$ for frame i [18].

Speech utterances for model training were recorded by three talkers, of combined duration 100 s, under the same conditions as those used in the experiments presented in Section 5. They were labelled as $\{V,U,S\}$ by hand then excluded from use in the subjective test set. The maximum likelihood estimate of the mean vector $\mathbf{m}_\omega$ and the covariance matrix $\mathbf{\Sigma}_\omega$ was determined for each class $\omega \in \{V,U,S\}$ and the relative frequency of each class was used to determine the prior probabilities $P\{\omega\}$. The probability of a feature vector $\mathbf{x}_i$ belonging to class ω is determined using Bayes's rule,

$$P\{\omega|\mathbf{x}_i\} = \frac{P\{\omega\} f_X(\mathbf{x}_i|\omega)}{f_X(\mathbf{x}_i)} \tag{2}$$

where $f_X(\mathbf{x}_i|\omega)$ is the class likelihood (determined by $\mathbf{m}_\omega$ and $\mathbf{\Sigma}_\omega$) and the total likelihood is estimated using

$$f_X(\mathbf{x}_i) = \sum_\omega P\{\omega\} f_X(\mathbf{x}_i|\omega). \tag{3}$$

The classification of frame i may only depend on $\mathbf{x}_i$ in which case it is sufficient to use the numerator of (2), where the class is determined as $\max_\omega P\{\omega\} f_X(\mathbf{x}_i|\omega)$. When classification is a time-dependent process, as described in the following subsection, it is necessary to base the decision on $P\{\omega|\mathbf{x}_i\}$.

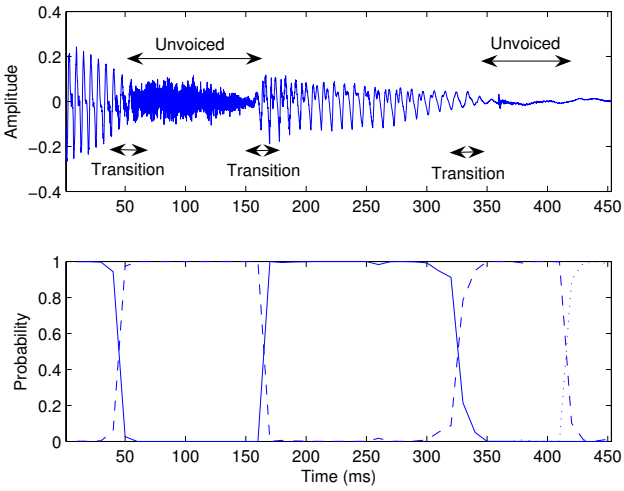

Figure 4: Normalised probabilities for $\omega \in \{V,U,S\}$. Solid: voiced, dashed: unvoiced, dotted: silence.

4.2 Determination of UT, Voiced and Silent Segments

The VUS detector provides a set of probabilities for voiced, unvoiced and silence as shown in Fig. 4. Voiced segments, $V(n)$, are identified by applying a Schmitt Trigger operator S^+ to the voiced probability, $S^+(P\{V|\mathbf{x}_i\})$. Transition segments, $T(n)$, are derived by identifying the boundaries of $V(n)$ and flagging a segment 10 ms before and after the boundary. Unvoiced segments, $U(n)$, are identified by applying a Schmitt Trigger to the unvoiced probability, $S^+(P\{U|\mathbf{x}_i\})$ (with upper and lower thresholds empirically set at $\{0.25, 0.75\}$) and extending in time scale by 2 ms at the boundaries. The UT segment is the union of $U(n)$ and $T(n)$, $UT(n) = U(n) \cup T(n)$ and all remaining segments are flagged as silence, $S(n)$.

5. SUBJECTIVE TESTING

A subjective test was performed to determine the mean opinions of the speech quality produced by two time scale modification algorithms. Both algorithms applied concatenative synthesis directly on speech recordings, using GCIs derived by the DYPSA algorithm. Algorithm 1 performed uniform time scale modification on the entire signal and Algorithm 2 is our proposed algorithm which uses UT detection to perform time scale modification during voiced speech and silence only.

The recording apparatus comprised an AKG C480 microphone connected to an RME Fireface 800 audio interface. Subjective testing samples were played back through the same interface connected to a pair of Sennheiser HD650 headphones.

Three talkers (two male, one female) were placed in an anechoic chamber and recorded speaking five phonetically balanced sentences at what they considered to be a 'normal' speaking rate and a 'fast' speaking rate (approximately 0.5-0.75 the duration of normal). The texts were taken from the APLAWD and TIMIT databases [19, 20]. The recorded speech was free from background noise, reverberation or any significant distortion.

An *ITU-T P800* [21] double-blind controlled test was employed where 30 test subjects were each played 60 random combinations of talker {1-3}, sentence {1-5}, talking rate {*normal*, *fast*}, algorithm [*no UT detection*, *UT detection*] and time scale modification factor {0.25, 0.5,..., 2.75, 3}. A modification factor of 1 implies no processing was undertaken and the subjects were not aware of what samples they were listening to, nor were they aware of how many algorithms had been employed. Sentences recorded 'fast' only had time-scale expansion, so that the modification factor was always greater than 1. The test subjects were asked to give overall opinion scores in the range {1-5}, paying attention to intelligibility, prosody

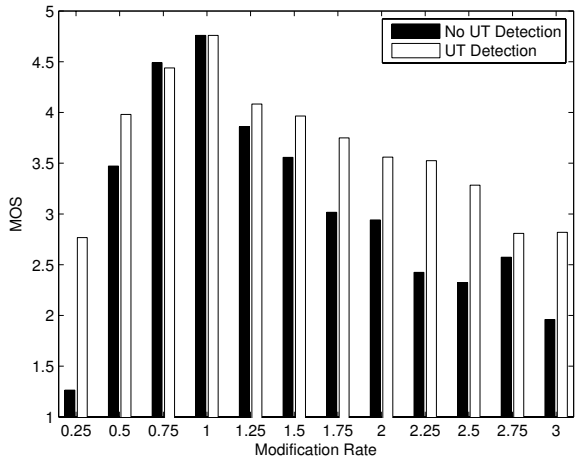

Figure 5: Mean Opinion Scores as a function of modification factor for a normal original talking rate.

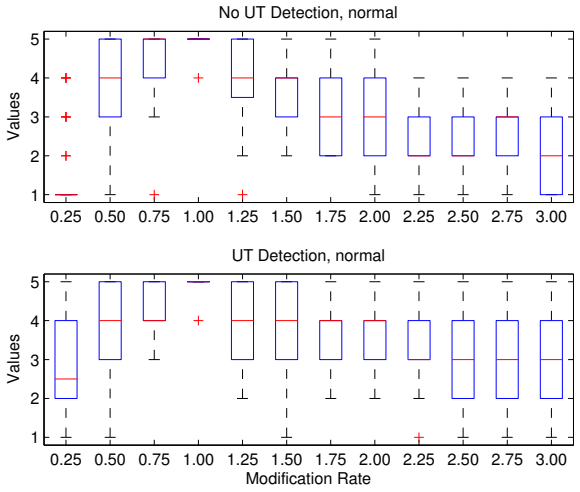

Figure 6: Confidence intervals of Mean Opinion Scores as a function of modification factor for a normal original talking rate. Boxes show the median, upper and lower quartiles, whiskers the most extreme values within 1.5 times the interquartile range and + the outliers.

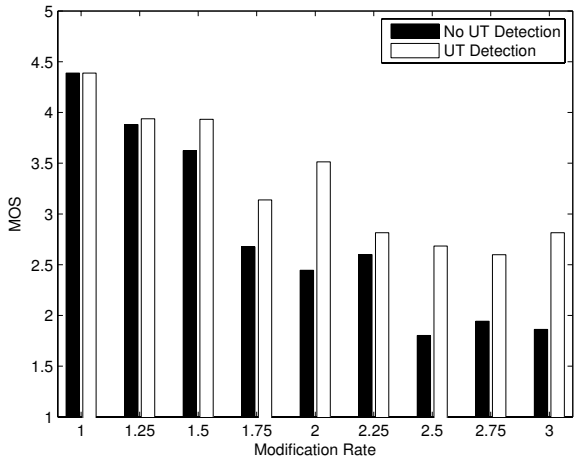

Figure 7: Mean Opinion Scores as a function of modification factor for a fast original talking rate.

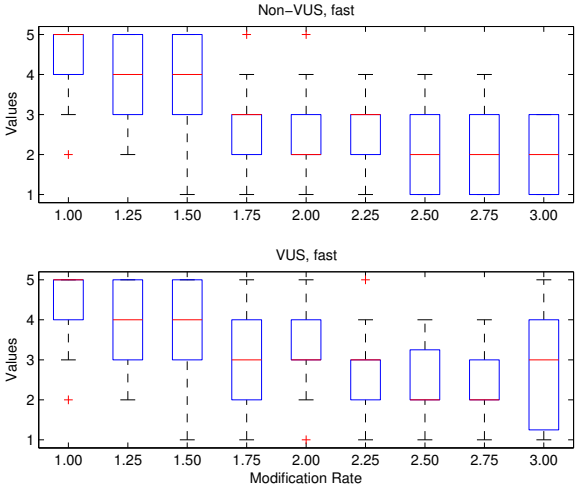

Figure 8: Confidence intervals of Mean Opinion Scores as a function of modification factor for a fast original talking rate. Boxes show the median, upper and lower quartiles, whiskers the most extreme values within 1.5 times the interquartile range and + the outliers.

and artefacts. Calibrated examples were given before the test was undertaken, defined as: 1=Unsatisfactory, 2=Poor, 3=Fair, 4=Good, 5=Excellent.

5.1 Results and Discussion

Figs. 5 and 7 show Mean Opinion Scores (MOSs) as a function of time scale factor with their corresponding confidence intervals in Figs. 6 and 8. Mean MOS scores are shown in table 1.

The results show that UT detection is preferred by the listeners, particularly at larger modification factors. However, at the lowest increments, 0.75 and 1.25, there may be evidence to suggest that UT detection is unnecessary. Informal listening has shown that although artefacts are reduced in the UT case, the flow is slightly interrupted so there may be a preference for smooth flow over artefacts for low levels of modification.

The greatest difference in opinions occurs at a factor of 0.25 on normal speech, where some subjects described the non-UT method as 'garbled' and the UT as 'unnatural but intelligible' during informal listenings. This would suggest that in the case of extreme speeding up, a listener prefers to preserve intelligibility at the cost of impairing natural flow.

The control samples show the highest MOS, though it is reduced by about 0.5 for 'fast' talking rate compared with 'normal'; similar scores are seen for normal speech modified by 0.5-0.75. This is evidence that intelligibility is preferred over the presence of artefacts at large deviations from normal; if this were not the case then the MOS for unmodified speech would be similar regardless of the original talking rate.

Now that it has been established that segmented time scale modification is a worthwhile pursuit, a possible extension of this method is the discrimination between different types of speech in addition to UT detection, then applying different amounts of stretching or compression based upon training data. This is mentioned in [8] where vowels are detected as a subset of voiced speech, but other cases such as inter-phoneme and inter-word pauses, stressed phonemes etc. may also vary differently with talking rate in natural speech.

6. CONCLUSION

Time scale modification is the altering of the length of a speech segment without changing pitch, prosody or formant structure. A

Table 1: Mean MOS scores

	Normal Speed	Fast Speed	Mean Speed
UT detection, μ_{UT}	3.5437	3.1792	3.3614
No UT detection, μ_{NUT}	2.8985	2.6044	2.7515
$\mu_{UT} - \mu_{NUT}$	0.6452	0.5748	0.6099

method for time scale modification has been proposed that gives good perceptual quality for a range of modification factors as demonstrated by subjective testing. The method employs an unvoiced / transition (UT) detector which ensures that time scale modification is only applied to silence or voiced segments, avoiding the artefacts caused by the time scale modification of unvoiced and transition sounds. GCIs are provided by the DYPSA algorithm as part of a practical, fast, reliable approach for time scale modification.

The method was tested by MOS testing against the same approach but excluding the UT detector. The results suggest that UT detection is preferred, though more so at larger modification factors. At small modification factors the benefit of UT detection is less pronounced.

The ability to alter the rate of speech reliably can enable applications such as time compression for the fast scanning of recorded messages, expansion for improving intelligibility or a mixture such as in the case of audio-video synchronization.

REFERENCES

[1] M. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Trans Acoustics, Speech and Signal Processing*, vol. 29, no. 3, pp. 374–390, Jun 1981.

[2] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–467, Dec. 1990.

[3] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, pp. 175–205, Feb. 1995.

[4] J. Makhoul, "Linear Prediction: A tutorial review," *Proc IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[5] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans Audio, Speech, Language Porcessing*, vol. 14, no. 3, pp. 972–980, May 2006.

[6] W. Verhelst, W. Verhelst, and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP-93*, M. Roelands, Ed., 1993, vol. 2, pp. 554–557 vol.2.

[7] J. di Martino and Y. Laprie, "Suppression of phasiness for time-scale modifications of speech signals based on a shape invariance property," in *Proc IEEE Intl Conf Acoustics, Speech and Signal Processing*, 7-11 May 2001, vol. 2, pp. 853–856.

[8] O. Donnellan, E. Jung, and E. Coyle, "Speech-adaptive time-scale modification for computer assisted language-learning," in *Advanced Learning Technologies, 2003. Proceedings. The 3rd IEEE International Conference on*, 9-11 July 2003, pp. 165–169.

[9] Sungjoo Lee, Hee Dong Kim, and Hyung Soon Kim, "Variable time-scale modification of speech using transient information," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, 21-24 April 1997, vol. 2, pp. 1319–1322vol.2.

[10] M. Demol, W. Verhelst, K. Struyve, and P. Verhoeve, "Efficient Non-Uniform Time-Scaling of Speech with WSOLA," in *Proc. Speech and Computers*, Patras, Greece, October 17-19 2005, pp. 163–166.

[11] D. Kapilow, Y. Stylianou, and J. Schroeter, "Detection of Non-Stationarity in Speech Signals and its Application to Time-Scaling," in *Proc European Conf on Speech Communication and Technology*, Budapest, Hungary, September 5-9 1999, pp. 2307–2310.

[12] M. Covell, M. Covell, M. Withgott, and M. Slaney, "MACH1: nonuniform time-scale modification of speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, M. Withgott, Ed., 1998, vol. 1, pp. 349–352 vol.1.

[13] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of Glottal Closure Instants in Voiced Speech using the DYPSA Algorithm," *IEEE Trans. Speech Audio Processing*, vol. 15, no. 1, pp. 34–43, January 2007.

[14] T. Ebihara, Y. Ishikawa, Y. Kisuki, T. Sakamoto, and T. Hase, "Speech synthesis software with variable speaking rate and its implementation on a 32-bit microprocessor," in *Consumer Electronics, 2000. ICCE. 2000 Digest of Technical Papers. International Conference on*, 13-15 June 2000, pp. 254–255.

[15] B. Yegnanarayana and R. Smits, "A robust method for determining instants of major excitations in voiced speech," in *Proc IEEE Intl Conf Acoustics, Speech and Signal Processing*, May 1995, pp. 776–779.

[16] Patrick A. Naylor, Jingjing Cui, and Mike Brookes, "Adaptive Algorithms for Sparse Echo Cancellation," *Signal Processing*, 2006, to appear.

[17] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans Acoustics, Speech and Signal Processing*, vol. 24, no. 3, pp. 201–212, June 1976.

[18] S. Furui, "Comparison of Speaker Recognition Methods using Statistical Features and Dynamic Features," *IEEE Trans Acoustics, Speech and Signal Processing*, vol. 29, pp. 342–350, June 1981.

[19] G. Lindsey, A. Breen, and S. Nevard, "SPAR'S Archivable Actual-Word Databases," Technical report, University College London, June 1987.

[20] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," *Proc. DARPA Workshop on Speech Recognition*, pp. 93–99, Feb. 1986.

[21] ITU-T, "Methods for subjective determination of transmission quality," ITU-T Recommendation P.800, Aug. 1996.