Time-scale modification as a speech therapy tool for children with verbal apraxia

E Coyle[1], O Donnellan[2], E Jung[3], M Meinardi[4], D Campbell[5], C MacDonaill[6] and P K Leung[7]

[1,2,3]School of Control Systems and Electrical Engineering,
[4,5]School of Languages,
Dublin Institute of Technology, Kevin Street, Dublin 8, IRELAND

[6,7]Digital Media Centre, Dublin Institute of Technology,
Aungier Street, Dublin 8, IRELAND

eugene.coyle@dit.ie, *olivia.donnellan@dit.ie*

[1,2]*www.dmc.dit.ie/ditcall*

ABSTRACT

A common suggested treatment for verbal apraxia is repetition, and the use of slow speech. The required slow speech may be attained by time-scaling ordinary-speed speech. However, when used for this purpose, the quality of the expanded speech must be of a very high quality to be of pedagogical benefit. This paper describes a new method of time-scaling based on the knowledge of speech characteristics, the relative durations of speech segments, and the variation of these durations with speaking rate. The new method achieves a high quality output making it suitable for use as a computer-assisted speech therapy tool.

1. INTRODUCTION

Verbal apraxia is a motor speech disorder. It manifests itself as an inability to consistently position the articulators for the production of speech sounds, and for sequencing those sounds into syllables or words. It is not a muscle or a cognitive disorder. Generally, the sufferer has a much better understanding of language than production of language. He/she may have the concept of the words, but has difficulty in translating from the planning area of the brain to actual words or speech.

Research into treatments for apraxia is still at an early stage, and approaches differ to some extent, however, there are some common features. The common themes among suggested therapies include a high degree of practice and repetition, and the use of slow speech. Experienced therapists report that children with apraxia need frequent repetition of sounds, sound sequences, and movement patterns in order to incorporate them and make them automatic. Indeed, a CD, Moir (2000), has already been produced for this purpose. The CD consists of a number of popular children's songs, but recorded at a slower rate, so that children with speech difficulties such as apraxia can have fun singing along, while at the same time developing their speech motor skills.

Our proposal is a variation of this concept, where poems and nursery rhymes are slowed down to any desired speed using a high-quality time-scale modification algorithm. The advantage of this is that the learner can choose a speed to suit his/her level, and can speak along with or repeat any chosen segments of the slowed-down clips at the chosen suitable speed. As the child becomes more adept, the speed can be increased, until eventually the learner will be able to repeat the segments at full speed. This provides the children with access to frequent repetition of sounds and sound sequences necessary to develop their motor skills, at a speed appropriate to the particular child's level of difficulty.

Time scale modification (TSM) refers to the process of altering the duration of an audio segment. A signal may be expanded, producing a signal of longer duration (a slower signal), or compressed, resulting in a signal of shorter duration (a faster signal). To be effectively time-scaled, the modified signal must retain all the characteristics of the original signal. In particular, the perceived pitch, speaker identity and naturalness must be maintained. Simply adjusting the playback rate of the signal will alter the duration of the signal, but

will also undesirably affect the frequency contents. Of the current methods for performing TSM, many are capable of producing a good quality output. However, for the proposed Computer-Assisted Speech Therapy (CAST) tool the quality needs to be extremely good, void of any distortion or unnaturalness. Current techniques simply do not produce a sufficient degree of quality, and furthermore, the quality decreases as the scaling rate increases.

Section 2 of this paper describes the general principles of a computationally efficient time-scaling technique and discusses some of the problems time domain over-lap add (TDOLA) methods encounter. Section 3 shows how to overcome these problems by suitably pre-processing the speech signal by taking into consideration specific characteristics of natural speech. This leads to a new speech-adaptive time-scaling algorithm, which provides a high quality output, even for high modification factors.

2. TIME-SCALE MODIFICATION

There are many different time-scaling techniques already developed, such as time-domain overlap-add techniques (TDOLA), frequency-domain techniques, and parametric techniques, with advantages and disadvantages to each. A full summary of these can be found in Lawlor (1999). The TDOLA approach is best suited to periodic signals such as voiced speech, and is also the best compromise of quality and efficiency, providing a high quality output for a relatively low computational load.

A TDOLA technique, in basic terms, performs time scale modification essentially by duplicating small sections of the original signal, and adding these duplicated segments using a weighting function, i.e., to make a signal have a longer duration, individual small segments are made longer my duplicating or repeating them. The technique requires firstly segmenting the waveform into a series of overlapping frames by windowing the signal at intervals along the waveform. These frames can then be added together, but with a different amount of overlap. Time-scale expansion is achieved by creating a waveform through the recombination of frames with a reduced amount of overlap. The amount of overlap required depends upon the desired expansion. Similarly, an increased amount of overlap will result in a time-scale compressed signal. The different TDOLA techniques generally vary in the way the waveform is segmented (choice of window, where segmentation occurs etc), or how successive frames are overlapped.

The most commercially popular TDOLA algorithm is the Synchronised Overlap-Add (SOLA) algorithm, Roucus (1985), because of its low computational burden with relatively high quality output. A more recent development, by Lawlor (1999), is the Adaptive Overlap-Add (AOLA), which offers an order of magnitude saving in computational burden without compromising the output quality, making it a suitable candidate for real-time implementation, such as in a computer-assisted speech therapy (CAST) package.

2.1 The Adaptive Overlap-Add Algorithm

The AOLA algorithm works in the following manner:

- A window length of ω is chosen such that the lowest frequency component of the signal will have at least two cycles within each window.
- The frame is duplicated.
- The duplicate of the original is shifted to the right to align the peaks, figure 1(b).
- Overlap-adding the original frame and its duplicate produces a naturally expanded waveform; figure 1(c). The length of this expanded segment is $\omega.ne$, where *ne* is the natural expansion factor.
- A portion of length *st* of the input signal is taken and is concatenated with the last expanded segment; figure 1 (d)-(e). *st* varies for each iteration and is a function of ω, *ne* and *de* (desired expansion factor).

$$st = \omega \frac{(1-ne)}{(1-de)}$$

- The next segment to be analysed is the ω-length frame ending at the right edge of the appended *st* segment, figure 1(e). This process continues until the end of the input signal is reached.

This method has a low computational load relative to other commercial algorithms of similar quality. Another advantage is that there are no discontinuities at the frame boundaries, as can be the case in other algorithms. This is because, referring again to figure 1, the area in (c) ending in the vertical dashed line and the area ending in the vertical dashed line in (d) are exactly the same shape, so the segment *st* appended to the expanded waveform will be aligned perfectly (e).

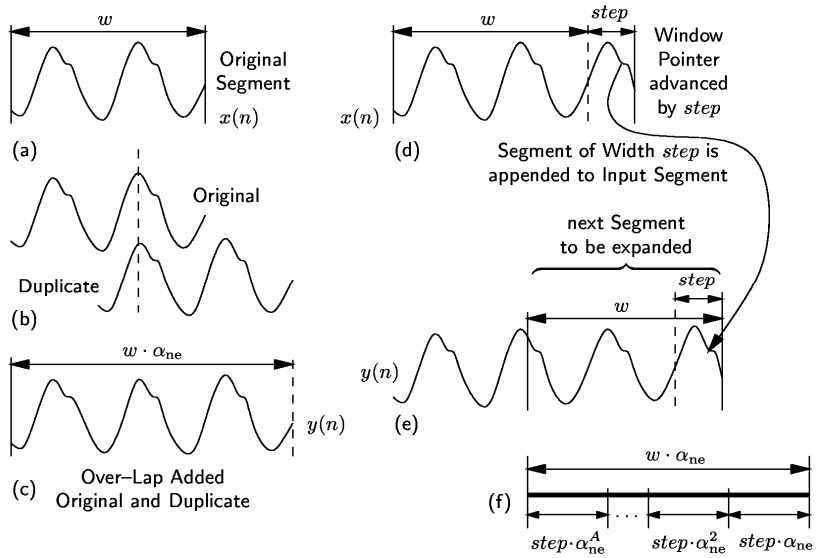

Figure 1. *AOLA algorithm, Lawlor (1999)*

2.2 Problems with Overlap-Add Techniques

Although TDOLA techniques provide the best compromise of computational load and quality, they have a few problems. As previously mentioned, TDOLA techniques perform time scale expansion by duplicating or repeating small segments of the original signal. If one of these segments to be repeated consists of a transient, such as a plosive, this may result in unwanted clicks, or what could be perceived as a 'stuttering' effect. Current TDOLA techniques assume that all segments of the speech, whether voiced, unvoiced, vowel or consonant, should be time-scaled at a uniform rate. A problem with this is that the transients (e.g. plosives in speech or drumbeats in music) are time expanded to the same degree as non-transient segments of the original signal. If a plosive were expanded using the TDOLA technique, the sound of the plosive would be distorted, and intelligibility of the resulting speech will be diminished. Also, vowels tend to be more influenced by speaking rate the consonants. To maintain intelligibility and naturalness, different time scaling factors need to be applied to the different segments of speech.

3. SPEECH-ADAPTIVE TIME-SCALE MODIFICATION

The intention of the proposed time-scaling algorithm is to be able to slow down speech while maintaining the naturalness and retaining all the characteristics of the original signal. To be capable of doing this certain aspects about the nature of the speech segments, the relative durations of these segments and the variation of these durations with speaking rate are considered for the time-scaling.

3.1 Relative variances in durations of vowels, voiced consonants and unvoiced consonants

It has been noted by Ebihara (2000) that the duration of unvoiced segments of human speech varies less than the duration of voiced segments. Ebihara recommends that a non-uniform rate be applied when time-scaling speech, so as to maintain the temporal structure of the utterance. He proposes a method of modifying only the voiced segments or alternatively only the vowel segments. Kuwabara (1997) backs up this theory by pointing out that changes in duration due to speaking rate are most obvious in voiced segments and particularly in vowels. He claims that the duration of voiced consonants varies more strongly with variations in speaking rate than that of unvoiced consonants.

By this reasoning, it can be concluded that, to imitate real speech characteristics, vowel sounds need to be more affected by time-scaling than consonants, and voiced consonants more than that of unvoiced consonants.

3.2 Consistency in duration and character of plosives

As previously mentioned, the effect of time-scaling on plosives is undesirable. As plosives convey a large amount of information, it is necessary to preserve their character under TSM. At large TSM factors, plosives may be artificially transformed into fricatives, e.g. /p/ slowed down at a high TSM factor may sound more like the fricative /f/. In normal speech, the closure stage of a plosive tends to be consistent in duration,

regardless of the speed of the speech. The duration of the burst also tends to be constant and the 'suddenness' of the onset of energy needs to be maintained. Time-scaling the burst leads to transient repetition, therefore to maintain the character of plosives the closure and burst need to be directly translated to the output without applying TSM.

3.3 Algorithm flowchart and summary

The following diagram (Figure 2) summarises the procedure proposed to achieve speech adaptive time-scale modification. Firstly, each segment of the input signal is examined to verify that speech exists. If no speech exists, the segment is assumed to be silence, for example a pause between phrases or sentences, and this segment of silence will then be time-scaled with a scaling factor of α_1. If speech exists, the type of speech contained in the segment must be determined. If analysis of the segment reveals that it is a plosive, or part of a plosive (closure or burst), the segment is copied to the output without any time-scaling, so as to preserve the nature of the plosive. If the segment is not a plosive, then it contains speech that is either voiced or unvoiced. Voiced speech is further analysed to determine whether it is a vowel or voiced consonant. As vowels are most influenced by speaking rate, they are time-scaled the most, with a scaling factor of α_1. The duration of voiced consonants varies less than vowels, but more than unvoiced consonants, so voiced consonants are time-scaled with a factor of α_2 and unvoiced speech is time-scaled with a scaling factor of α_3, where

$$\alpha_1 > \alpha_2 > \alpha_3 > 1$$

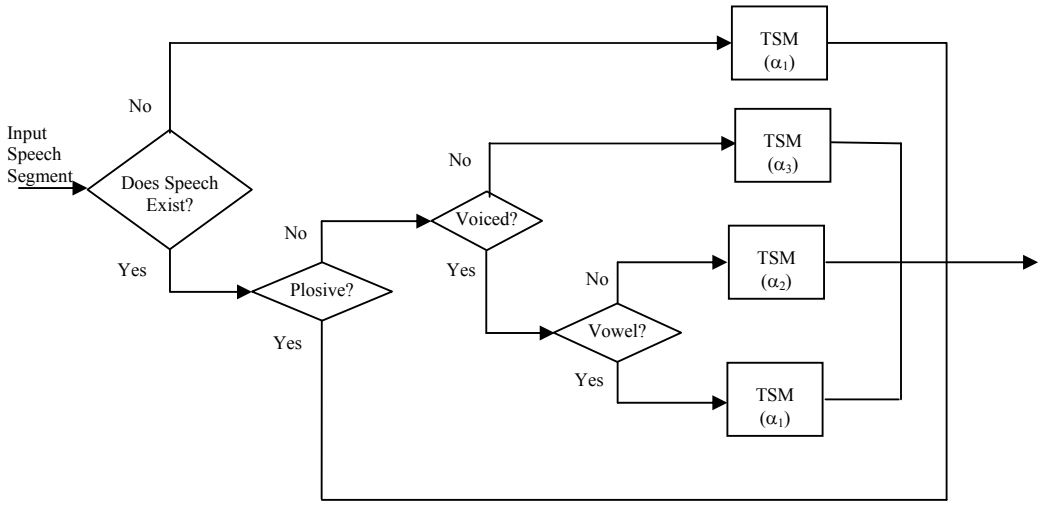

Figure 2. *Flowchart for the proposed speech adaptive time-scaling method*

4. EXPERIMENTS AND RESULTS

A number of speech samples were recorded at a sampling rate of 16 kHz, and more samples were taken from the TIMIT database. An equal number of male and female speakers were used. Each signal was then analysed on a frame-by-frame basis, and manual segmental detection was performed, i.e. through a combination of examining the waveform and listening to the signal, decisions were made as to whether the frame consisted of a plosive, vowel, voiced consonant, unvoiced consonant or silence. Matlab scripts were then written to adapt the AOLA algorithm and perform a variety of variable and uniform time-scaling methods, as described in Table 1.

For the evaluation of the performance of the different time-scaling methods a series of informal listening test was conducted. Two speech samples were recorded at a sampling rate of 16 kHz and two more samples were taken from the TIMIT database. Seven male and seven female listeners participated in the test. Each test signal was segmented depending on the utterance type (plosive, vowel, voiced consonant, unvoiced consonant or silence) and then slowed down using the methods described in Table 1. All slowing down methods were based on the AOLA algorithm and the implementation was done in Matlab.

For method D, the new proposed method, two different sets of scaling parameters were considered in order to investigate the existence of a difference in quality for different sets of parameters. For each set, the

requirement of $1 < \alpha_3 < \alpha_2 < \alpha_1$ was adhered to, but the distance between the values was varied. In the first set (D1), the values varied linearly from 1 to α_1, while in the second set (D2), these values varied exponentially.

Table 1. *Compared time-scaling methods*

Method	Description
A	Uniform TSM
B	Variable TSM with voiced segments only being modified
C	Variable TSM with vowels only being modified
D	The new speech-adaptive TSM method.

4.1 Experimental Setting

All speech samples were time-scaled by each of the above methods and at three different overall time-scale modification factors, namely 2, 2.5 and 3. Two different informal listening tests were used to assess the quality of the techniques. The first consisted of 12 preference tests, in which all methods were compared. For each test, the subjects were asked to rank 5 different tracks, each of which contained a speech signal time-scaled using one of the methods A, B, C, D1 or D2. The second part consisted of eight pair comparisons, in which the proposed method (D) was compared to a traditional plain uniform-scaling method (A).

4.2 Test results

The results of the experiments show a clear preference for the proposed method, with 88% of listeners choosing a signal time-scaled by this method as their first choice in part one of the tests (Table 2).

Table 2. *First preference allocations*

Method:	First Preferences
D	88%
A or B or C	12%

The outcome of the overall rankings show a small improvement in quality of methods B and C compared to that of A, but methods D1 and D2 lead the field by a much more significant amount (Figure 3). This pattern is noticeable for all time-scaling factors investigated, as can be seen in Table 3.

Table 3. *Preference test results for different overall scaling factors*

	SF 2	SF 2.5	SF 3
First	D2	D2	D2
Second	D1	D1	D1
Third	A	C	C
Fourth	C	B	B
Fifth	B	A	A

Also evident from Table 3 is the deterioration in quality of method A as the time-scaling factor increases. This can be observed more clearly from the results of the second part of the test, in which 78% of listeners chose method D over method A (Table 4). The variation in this value with scaling factor forms the interesting result that, whereas method A decreases in quality as the scaling factor is increased, method D maintains a high quality output, as seen in Figure 4.

5. FUTURE WORK

Currently, various different techniques are being examined to determine the most efficient way of automatically detecting the different segmentals. Listening tests will be conducted to determine the optimum TSM factors to be used for each type of segmental. Real-time systems have been created within the research team, with implementation of the developed digital signal processing TSM algorithms, and a future task will be the development of a user-interface to enable the program to be used as a computer assisted speech therapy tool. This prototype can then be evaluated and tested by speech therapists and children with apraxia, and for this trial product, work is commencing on time-scaling of nursery-rhymes and the associated issues involved. Another future task will be to investigate the use of time-scale modification as an assistive-technology for other disabilities, such as stuttering, teaching pronunciation to the visually impaired, and in the rehabilitation of stroke victims.

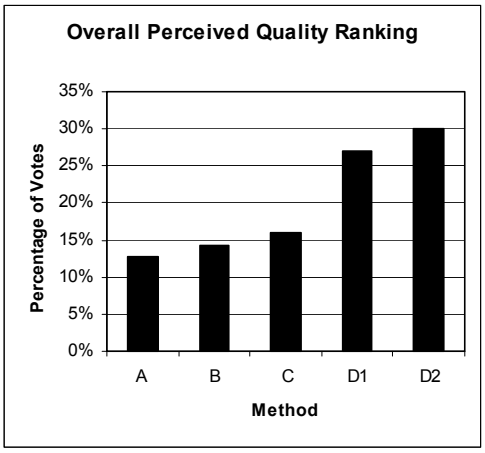

Figure 3. *Preference test results.*

Table 4. *Comparison preferences.*

Method:	Preferences
D	78%
A	22%

6. CONCLUSIONS

This paper discusses the merits of slowing down speech samples for use in a new computer-based speech therapy tool that allows poems and nursery rhymes to be slowed to any desired speed with minimal affect on the output quality. To be of benefit for this purpose the quality of the time-expanded speech needs to be very high. Several techniques to achieve high quality slowed speech were described and a new method using adaptive speech scaling was introduced. The tests carried out compared the quality of four different slow-down methods, namely uniform scaling, scaling only voiced segments, scaling only segments containing vowels and finally applying three different scaling factors to various types of speech segments while not scaling segments containing plosives.

The listening test results show that the proposed method using full segmental distinction is superior to the other methods and clearly delivers the best results. When large scaling factors are applied the advantage of the adaptive speech scaling method becomes even more apparent over traditional uniform speech scaling. The best results were achieved when the distance between the three different scaling factors increased exponentially. The proposed speech-adaptive slow-down system is therefore the most beneficial for the application in a CAST system.

Acknowledgements: This work was funded by the Enterprise Ireland administered Advanced Technologies Research Programme (ATRP) 2001, - Project ATRP/01/203, "DITCALL – Digital Interactive Tools for Computer Assisted Language Learning."

7. REFERENCES

T Ebihara, Y Ishikawa, Y Kisuki, T Sakamoto, T Hase (2000), Speech Synthesis Software with Variable Speaking Rate and its Implementation on a 32-bit Microprocessor, *19th IEEE International Conference on Consumer Electronics (ICCE 2000)*, Los Angeles Airport Marriott, USA

H Kuwabara (1997), Acoustic and Perceptual Properties of Phonemes in Continuous Speech as a Function of Speaking Rate, *Proc. Eurospeech 97*, pp. 1003-1006.

B Lawlor (1999), *Audio Time-Scale and Frequency-Scale Modification*, PhD Thesis, Department of Electrical Engineering, University College Dublin.

R Moir, M Sturm (2000), Time to Sing! CD, http://www.time2sing.

S Roucus, A M Wilgus (1985), High-Quality Time-Scale Modification for Speech, *IEEE Proceedings on Acoustics, Speech and Signal Processing,* pp. 493-496.