2003-01-01

Speech-adaptive time-scale modification for computer assisted language learning

Olivia Donnellan
Dublin Institute of Technology, olivia.donnellan@dit.ie

Elmar Jung
Dublin Institute of Technology, elmar.juung@dit.ie

Eugene Coyle
Dublin Institute of Technology, Eugene.Coyle@dit.ie

Speech-adaptive time-scale modification for computer assisted language learning

Olivia Donnellan* Elmar Jung†
Eugene Coyle‡

*Dublin Institute of Technology, olivia.donnellan@dit.ie
†Dublin Institute of Technology, elmar.jung@dit.ie
‡Dublin Institute of Technology, Eugene.Coyle@dit.ie

— Use Licence —

Attribution-NonCommercial-ShareAlike 1.0

You are free:

- to copy, distribute, display, and perform the work
- to make derivative works

Under the following conditions:

- Attribution.
  You must give the original author credit.
- Non-Commercial.
  You may not use this work for commercial purposes.
- Share Alike.
  If you alter, transform, or build upon this work, you may distribute the resulting work only under a license identical to this one.

For any reuse or distribution, you must make clear to others the license terms of this work. Any of these conditions can be waived if you get permission from the author.

Your fair use and other rights are in no way affected by the above.

Speech-Adaptive Time-Scale Modification for Computer Assisted Language-Learning

Olivia Donnellan
Dublin Institute of Technology
E-mail: olivia.donnellan@dit.ie

Elmar Jung
Dublin Institute of Technology
E-mail: elmar.jung@dit.ie

Eugene Coyle
Dublin Institute of Technology
E-mail: eugene.coyle@dit.ie

Abstract

In traditional foreign language learning programmes students are offered a tutor model characterised by slow, deliberate speech. This is insufficient to prepare them to cope with native, colloquial speech. By applying a time-scale modification (TSM) algorithm to natural-speed, native speech, students have access to a more desirable, natural speech corpus which permits them to practise essential listening skills in a more focussed manner. This paper presents a method which allows slowing down speech without compromising the quality, pitch or naturalness of the slowed speech by applying different scaling factors to different types of speech segments. The new method is compared to traditional uniform-rate techniques and other variable time-scaling methods. The results show that the proposed approach produces a superior quality output, even for high modification rates.

1. Introduction

One of the major aims of Computer Assisted Language-Learning (CALL) packages is to enhance the student's perception and hence comprehension of the foreign language. Cauldwell [1] states, that "in the normal stream-like state of everyday spontaneous speech, the boundaries of words are blurred together…". The ability to slow down streamed speech in the acquisition of a second language improves the learner's essential perceptive skills for understanding what happens in natural speech. Hoven [2] shows, that a learner's ability to process spoken language is affected by the rate of delivery, and Brown [3] argues that colloquial language is the only practical model to be used in listening exercises.

Whereas slow, colloquial speech might be a suitable model for learners to imitate in their own speech production, it is an inadequate model for training learners to comprehend natural, native speaker (NS) speech flow – an essential task avoided by almost all speech tutors. Most teaching programmes use idealised recordings which tend towards a citation, or idealised form of speech with unnaturally clear diction. The phonological challenges facing the learner in natural speech by native speakers – which in everyday speech can be delivered at speeds of over 400 words per minute [1] – are never faced, and therefore the learner's improvement in comprehension skills must inevitably be slow.

Normal speed (i.e. fast!) NS to NS speech is characterised by phonemic reduction, elision and adaptation to the immediate phonological environment at segmental level. Stress patterns and suprasegmental features of NS-NS speech must be practised if aural comprehension is to be assured. Incorporation of a time-scale algorithm for slowing down speech into an oral/aural training package therefore permits a pedagogically more relevant approach denied to programmes lacking the facility.

By applying time-scaling sequentially, learners can slow down natural speech until these syntactically and lexically important features can be perceived, practised and mastered – at least for comprehension purposes, if not for actual speech production.

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 a CALL 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.

Proceedings of the The 3rd IEEE International Conference on Advanced Learning Technologies (ICALT'03)
0-7695-1967-9/03 $17.00 © 2003 **IEEE**

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 that provides a high quality output, even for high modification factors.

2. Time-domain overlap-add techniques

Many different TSM techniques have been developed, such as TDOLA techniques, frequency-domain techniques and parametric techniques. The advantages and disadvantages to each of them depend strongly on the application. The TDOLA approach is best suited to periodic signals such as 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 performs time-scale modification essentially by duplicating small sections of the original signal and appending these duplicated segments to the original segments after an alignment procedure thus building up the scaled signal. The different TDOLA techniques generally vary in the way the waveform is segmented, the choice of window, where segmentation occurs, etc., and how successive frames are aligned and overlapped. Especially the frame alignment procedure determines the quality and the computational performance of the algorithm.

2.1. The Adaptive Overlap-Add (AOLA) algorithm

The most commercially popular TDOLA algorithm is the Synchronised Overlap-Add (SOLA) algorithm [4], because of its low computational burden with relatively high quality output. A more recent development by Lawlor is the Adaptive Overlap-Add (AOLA) [5], which offers an order of magnitude saving in computational burden without compromising the output quality. This makes the method a suitable candidate for applications that require real-time performance, such as a CALL package. 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 and the chosen frame is duplicated
- The duplicate of the original is shifted to the right to align the peaks (Figure 1b).
- Overlap-adding the original frame and its duplicate produces a naturally expanded waveform (Figure 1c). The length of this expanded segment is $\omega \cdot \alpha_{ne}$, where α_{ne} is the natural expansion factor.
- A portion of length *step* of the input signal is taken and is concatenated with the last expanded segment; Figure 1 (d)-(e). *step* varies for each iteration and is a function of ω, α_{ne} and α_{de} (desired expansion factor):

$$step = \omega \frac{(1-\alpha_{ne})}{(1-\alpha_{de})} \qquad (1)$$

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

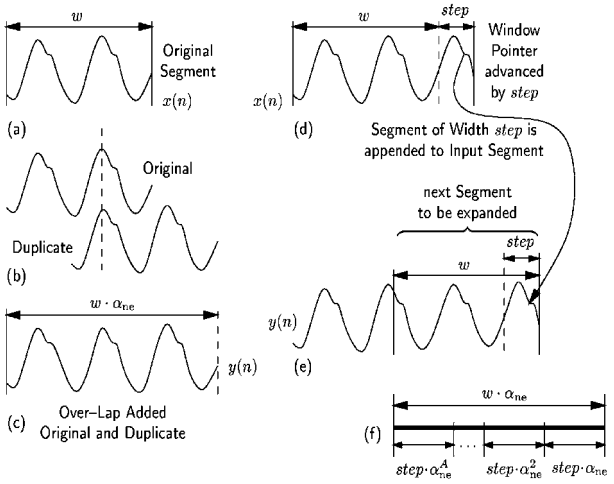

Figure 1. Principle of the AOLA algorithm [5]

The low computational load relative to other popular TSM algorithms of similar quality is not the only benefit of this method. 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 *step* appended to the expanded waveform will be aligned perfectly (e).

2.2. Problems with TDOLA techniques

Although TDOLA techniques provide the best compromise of computational load and quality they have a few problems affecting the output quality, especially when applied to speech. Current TDOLA techniques make no assumption of the nature of the signal they are applied to and generally apply a uniform scaling rate to all segments of the speech, whether voiced or unvoiced, or vowel or consonant. In real speech, some segments are more influenced by speaking rate than others. To maintain intelligibility and naturalness, different time

Proceedings of the The 3rd IEEE International Conference on Advanced Learning Technologies (ICALT'03)
0-7695-1967-9/03 $17.00 © 2003 **IEEE**

scaling factors need to be applied to the different segments of speech.

Another problem arising from uniform time-scaling 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. Plosives (/b/, /d/, /g/, /k/, /p/, /t/) are produced by complete closure of the oral passage and subsequent release with a burst of air. An example of a voiced plosive is the 'duh' sound in 'dog'; an unvoiced example is 'puh' as in 'pit'. There are three distinct stages in the production of a plosive: (a) closure, i.e. when the air-stream is totally blocked by the articulators and the air-pressure builds up behind the obstruction, (b) burst, i.e. a sudden increase in energy when the articulators quickly open and a burst of air rushes through and (c) transition, i.e. the transition segment to the following sound. If a plosive were scaled at the same rate as a vowel it would suffer distortion and thus intelligibility of the resulting speech would be diminished. Apart from a smearing effect a plosive can suffer from another distortion known as transient repetition, which can manifest itself as clicks or could be perceived as a 'stuttering' effect.

3. Speech adaptive time-scale modification

To overcome the described problems and limitations a more distinguished approach is proposed. 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 [6] 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 only the vowel segments. Kuwabara [7] 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. Proposed algorithm summary and flowchart

In the proposed algorithm firstly each segment of the input signal is examined to verify that speech exists. If no

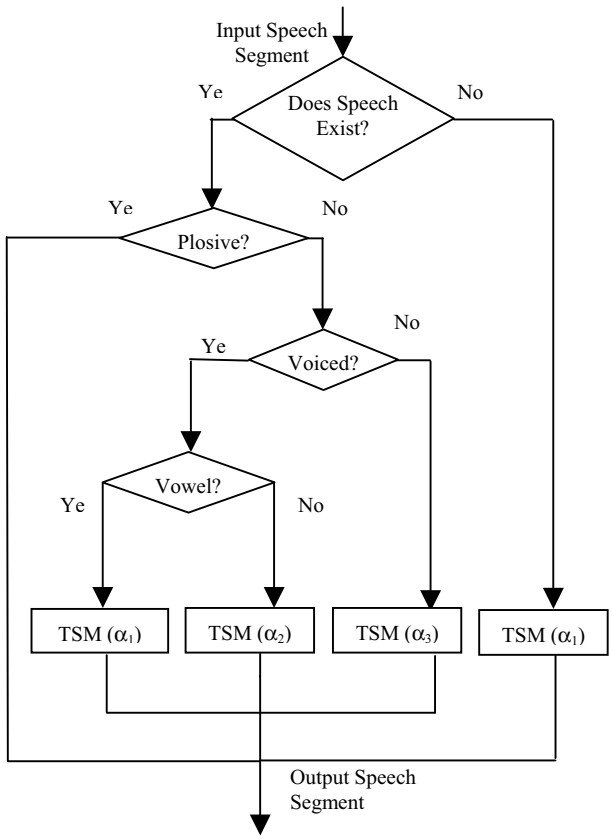

Figure 2. Flowchart for the proposed speech adaptive time-scaling method

speech exists, the segment is assumed to be silence, for example a pause between words 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 summarises the procedure proposed to achieve speech adaptive time-scale modification.

4. Experiments and results

For the evaluation of the performance of the different time-scaling methods a series of informal listening tests were 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.

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.

Table 1. Compared time-scaling methods

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.

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).

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

Table 2. First preference allocations

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).

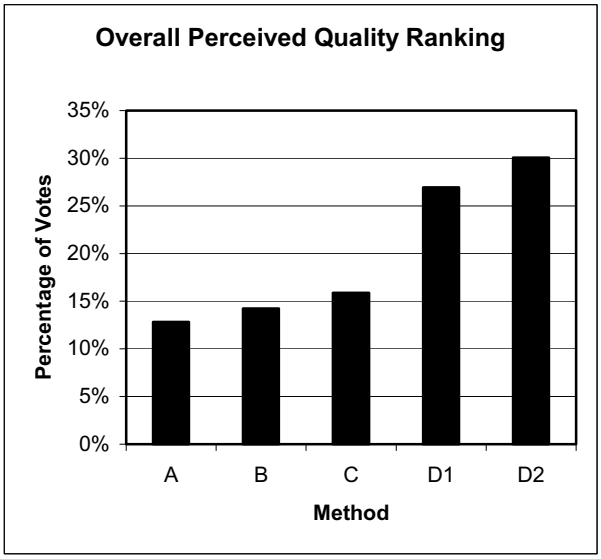

Figure 3. Preference test results

This pattern is noticeable for all time-scaling factors investigated, as can be seen in Table 3.

	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

Table 3. Preference test results for different overall scaling factors

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.

Method:	Preferences
D	78%
A	22%

Table 4. Comparison preferences

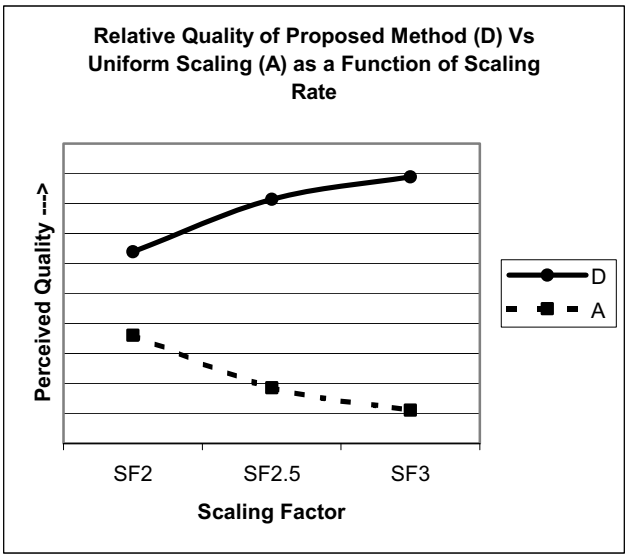

Figure 4. Relative quality of method D compared to method A as a function of scaling rate

5. Conclusion

This paper discusses the merits of slowing down speech samples for use in computer-assisted language-learning. To be of benefit for a language learner 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 voiced speech and 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 CALL system.

6. Acknowledgements

The support of the Informatics Research Initiative of Enterprise Ireland is gratefully acknowledged.

7. References

[1] Cauldwell, R., "Phonology for Listening: Relishing the Messy", *TESOL,* Salt Lake City, Utah, 2002.

[2] Hoven, D., "Towards a Cognitive Taxonomy of Listening Comprehension Tasks", *SGAV Review*, Vol. 9 No.2, 1991, pp. 1-12.

[3] Brown, G., *Listening to Spoken English*, 2nd Edition, Longman Pub. Co., 1991

[4] Roucus, S. and Wilgus, A.M., "High-Quality Time-Scale Modification for Speech", *IEEE Proceedings on Acoustics, Speech and Signal Processing*, March 1985, pp. 493-496.

[5] Lawlor, R., "A Novel Efficient Algorithm for Audio Time-Scale Modification", *Irish Signals and Systems Conference*, National University of Ireland, Galway, 1999.

[6] Ebihara, T., Ishikawa, Y., Kisuki, Y., Sakamoto, T. and Hase, T., "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, June 2000

[7] Kuwabara, H., "Acoustic and Perceptual Properties of Phonemes in Continuous Speech as a Function of Speaking Rate", *Proc. Eurospeech 97*, pp. 1003-1006.