Automatic Segmentation of Recorded Speech into Syllables for Speech Synthesis

Eric Lewis
Mark Tatham

Reproduced from: Lewis, E. and Tatham, M. (2001) Automatic segmentation of recorded speech into syllables for speech synthesis. *Proceedings of Eurospeech '01*, 1703-1707. Aalborg: International Speech Communication Association

Abstract

Concatenated waveform text-to-speech synthesis systems require an inventory of stored waveforms from which units of speech can be extracted for subsequent rearrangement and concatenation as needed. In previous papers [1], [2] we have argued that for natural sounding speech the syllable should be the preferred unit. The mark-up of the stored waveforms for segmentation into syllables must be precise and for our MeteoSPRUCE limited domain system the mark-up has been done by manual editing. In this paper we describe how most of the segmentation can be done automatically, leaving only those waveforms which would be prone to error to be segmented manually.

1. Introduction

MeteoSPRUCE is a limited domain concatenated waveform text-to-speech synthesis system based on the syllable as its fundamental unit of stored speech. It has an inventory consisting of recordings of 1725 monosyllabic and polysyllabic words. Those words for which recordings do not exist in the inventory are constructed by extracting words and/or syllables which *are* in the inventory and recombining them as appropriate. In [3] we provided rules on how syllables could be modified for concatenation in contexts other than those from which they were excised. Because MeteoSPRUCE is a limited domain system it was feasible to perform all the necessary mark-up of the waveforms by manual editing. To extend our synthesis system to unrestricted text it clearly becomes desirable to have procedures for the automatic segmentation of words into syllables. Moreover, we believe the future for text-to-speech speech synthesis systems resides in their being able to provide a variety of voices, Hence, while the normal effort in marking up one or two voices can be tolerated, the task clearly becomes intolerable for a significant number of voices.

2. Syllable definition

Before embarking on the task of automatic syllable detection one has to decide what constitutes a syllable in the first place. There is no universal agreement on a rigorous definition of the syllable but one which has wide acceptance, for English, is the following [4]:

- syllables can be expressed in the form $C^{3}_{4}VC^{4}_{3}$ where C^{n}_{0} signifies 0 to *n* consonants and V signifies a vowel.

A much wider discussion of what constitutes a syllable can be found in [5]. This definition still allows one some freedom in deciding where the syllabic boundaries should be because the definition is abstract and the marking of the syllables is done on the physical waveform. For example do we express windy as *win-dey* or *wind-ey*. We have decided that for engineering purposes a morphemic decomposition of words into syllables is to be preferred whenever possible with segmentation occurring on a phonological basis otherwise. The reason for this is that since syllables will be used for constructing new words then it is most likely that these words will be built-up on a morphemic basis.

With the above definition of a syllable, a syllable boundary can be one of the following types:

- V-V, V-C, C-V, C-C

Furthermore, since a consonant can be one of plosive, fricative, affricate, liquid or nasal there are, in theory, 36 rules relating to the classification of syllable boundaries. In practice we found that the number of rules is significantly reduced because in many cases they as so similar as to be not worthy of separate description.

3. Automatic syllable detection

As long ago as 1975 Mermelstein [6] proposed a technique for automatic segmentation of speech into syllabic units. His algorithm was based on using a loudness minimum as an indicator of syllabic boundaries. This technique was reasonably successful in achieving its aims and, indeed, our own algorithm makes use of a similar criterion in identifying possible boundaries. However, it cannot be used in isolation because the minima are phonetically determined and more closely allied to the phonological description of syllables rather than our morphemic definition.

The use of syllables as units in automatic speech recognition has gained in popularity over the last few years and a number of algorithms have been proposed for their automatic identification [7], [8], [9]. However, for speech recognition purposes the boundary of the syllable does not have to be precisely defined and so these algorithms are not so useful for speech synthesis.

The MeteoSPRUCE system uses a dictionary for its pronunciation phase, one of the advantages of such a system being that the syllabification of words is recorded in the dictionary. Additionally, the use of a TD-PSOLA algorithm for the imposition of the required intonation and rhythm onto the concatenated waveform requires precise marking of the periods. This can be done automatically with some subsequent manual editing for the resolution of the more difficult decisions as to whether a section of the waveform should be marked as voiced or not. The provision of both a phonetically defined syllable boundary together with precise knowledge of voiced and unvoiced sections of the waveform make the task of automatically marking the syllable boundaries in the waveform considerably easier. For example, the word *difficult* is listed in the dictionary as *~di-fi-kalt* and since the periods for this word have already been marked it is a simple matter to find the start of the two unvoiced sections to identify the beginning of the syllables *fi* and *kalt* as shown in Fig. 1. The notation used here is that known as the JSRU [10] notation with ~, ‘ and - indicating main stress, secondary stress and no stress respectively on their following syllables.

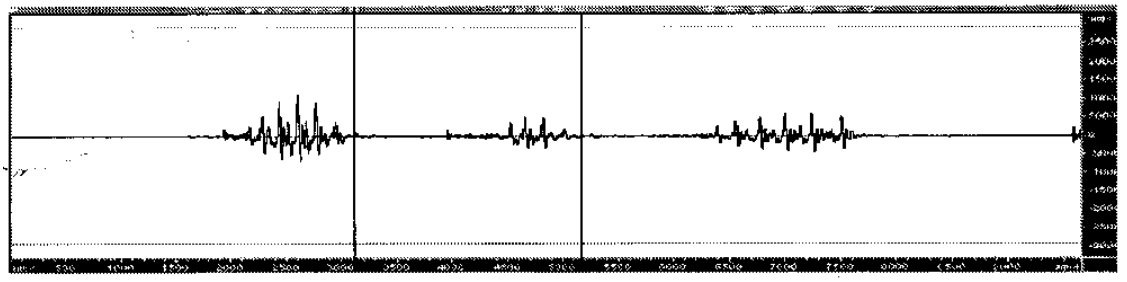

Figure 1: Syllable boundaries marked for the word *~di-fi-kalt*

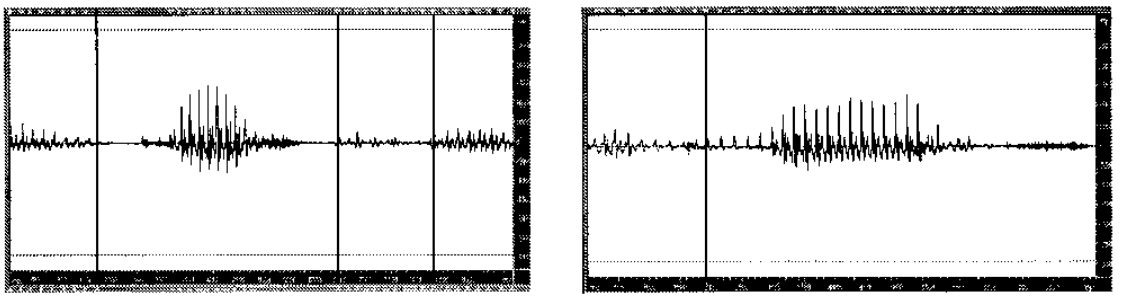

Figure 2: Syllable boundaries for the words *–un~plez-ant-ley* and *-ad~varns*

3.1. Rules for detection of syllable boundaries

The MeteoSPRUCE dictionary contains 1155 polysyllabic words providing 1730 examples of syllable boundaries of which 344 are different. On examination of these boundaries we found that it was not possible to give clearly defined rules in all cases. In the first instance we found that the behaviour of affricates was so similar to that of fricatives as to not warrant making them a separate category. Secondly, since automatically determining the precise boundary for v-v syllables is often very difficult and since their occurrence is relatively infrequent (less than 2% of the 1730 examples were V-V boundaries), we decided to refrain from the attempt. Instead syllables that need to be used in a V-V context for subsequent synthesis can be taken from another context and then modified to comply with their new context [3]. It also transpires that certain syllables can only be used in certain contexts and should not be transposed into contexts other than those in which they were recorded. For example, in a C-C combination where the consonants are both plosives, the first syllable does not contain the release of the plosive. Similarly, for a fricative-plosive boundary in which the fricative is an [s] and the plosive is unvoiced, the second syllable has its plosive sounding like its voiced counterpart after excision.

The initial search for syllable boundaries is based on locating local minima in the waveform amplitude. Although this will frequently detect more candidates for a syllable boundary than are required, we know exactly how many we need to retain because we know the number of syllables in the word. Furthermore, the information with regard to voicing, which is available from the pitch mark-up of the waveform, together with the knowledge of the boundary type obtainable from the dictionary, enables a clear decision to be made as to where precisely the syllable boundaries need to be located. Rules for the identification of syllable boundaries are given in Table 1 and two examples of words marked up according to these rules are given in Fig. 2.

On those occasions where the algorithm fails to determine the syllable boundaries it marks the word in the dictionary to indicate that no attempt should be made to use the individual syllables from that word. This will prevent these syllables and their context being listed in the syllable index.

4. Conclusions

Rules for the automatic segmentation of words into syllables have been derived based on a morphemic decomposition. An algorithm has been produced for the implementation of these rules which utilises the orthography-to-phoneme dictionary together with an accurate mark up of the pitch periods. Although not capable of doing a complete automatic segmentation of all words we believe that about 90% of the waveform mark up can now be automated allowing much faster derivation of synthetic voices.

5. References

[1] Lewis, E. and Tatham, M., "SPRUCE – a new text-to-speech synthesis system", *Proceedings of Eurospeech* '91. *ESCA, Genova,* pp 1235-1238, 1991.

[2] Tatham, M. and Lewis, E., "Syllable reconstruction in concatenated waveform speech synthesis", *Proceedings of the XIVth International Congress of Phonetic Sciences*, pp 2303-2306, 1999.

[3] Lewis, E. and Tatham, M., "Word and syllable concatenation in text-to-speech synthesis", *Proceedings of Eurospeech* '99, *ESCA, Budapest,* pp 615-618, 1999.

[4] Gimson, A.C., *An Introduction to the Pronunciation of English.* First Edition. Arnold, London, 1962.

[5] van der Hulst, H. and Ritter, N., (eds.) *The Syllable: Views and Facts*, Walter de Gruyter, Berlin, 1999.

[6] Mermelstein, P., "Automatic segmentation of speech into syllabic units", *J. Acoust. Soc. Amer.,* Vol. 58(4): 880-883, 1975.

[7] Green, P. D., Kew, N. R. and Miller, D. A, "Speech representations in the SYLK recognition project", *Visual Representation of Speech Signals*, Cooke, M. P. Beet, S. W. and Crawford M. D., editors, chapter 26, pp 265-272. John Wiley, 1993.

[8] Reichl, W. and Ruske, G., "Syllable segmentation of continuous speech with artificial neural networks", *Proceedings of Eurospeech* '93, *ESCA, Berlin.* pp 1771-1774, 1993.

[9] Wu, S.-L., Shire, M. L., Greenberg, S. and Morgan, N., "Integrating syllable boundary information into speech recognition", *Proceedings of the International Conference Acoustics, Speech and Signal Processing (ICASSP* '97), Vol. 2, 987-990, 1997.

[10] Lewis, E. "A 'C' implementation of the JSRU text-to-speech system", *Report TR-89-15*, Dept. Comp. Sci., Bristol University, England, 1989.

C-C syllable boundary		
Plosive – vowel Fricative – vowel Affricate – vowel	The second syllable begins with the onset of vocal cord vibration immediately following the release of the plosive. This behaviour is similar for all three cases. However, the frication is much longer for the case of a fricative and affricate rather than a plosive. Care needs to be taken when the consonant is voiced since voicing can occur throughout the boundary. In this case the start of the second syllable is marked by a jump in amplitude of the waveform.	'kon-ti~nent-al ~big-ist ~broak-en -a~kord-ing -in~krees-ing -un~plez-ant -a~taach-ing -kan~verj-ing
Liquid-vowel Nasal-vowel	Both the liquid and nasal are marked by a trough in the waveform, although it is normally longer for the nasal than the liquid. The nasal is also characterised by a drop in the spectral energy. The end of the trough is the start of the second syllable.	~ee-kwal-ing ~weir-a'bouts ~pre-val-ant ~kum-ing ~or-gan-iez ~gluum-i-ley
V-C syllable boundary		
Vowel - plosive	The syllable boundary is where the vocal cord vibration in the vowel of the first syllable drops sharply in amplitude. This is the start of the plosive stop phase. For unvoiced plosives the end of the vowel is easy to detect because there's a relatively long period of voiceless frication for the following plosive. For voiced plosives, however, it is more difficult because vocal cord vibration can continue for 20-30 ms into the start of the second syllable.	~ee-kwal ~noa-ba-dey ~kaa-pi-tal -pri~dik-shan ~dee-tail -bi~gin
Vowel – fricative Vowel – affricate	The behaviour for the vowel-fricative boundary is very similar to that for the vowel-plosive with voicing and frication overlapping for the voiced fricatives.	~di-fi-kalt -pra~vi-zhan ~nu-thing ~haa-za-das ~tem-pra-cha ~graa-ja-ley
Vowel – liquid	As the liquid is characterised by a trough the boundary is at the start of the trough. This can be difficult to determine but the point at which the magnitude of the signal in the trough is a minimum can be taken as the start of the second syllable. In this case the first syllable is not safe to use in another context.	-pri~li-mi-na-rey ~for-wad ~ei-rial -a~bai-yans
Vowel – nasal	Similar behaviour to that for the vowel-liquid boundary but the nasal trough is easier to recognise.	'aab~nor-mal 'kon-ti~nent-al

C-C syllable boundary		
Plosive – plosive	The syllable boundary can be taken as the middle of the fricative region formed from the two plosives. For this case the first syllable cannot be used in a different context because the first plosive is not usually released.	~kloud-berst 'sep~tem-ba ~faak-ta ~week'dai
Plosive – fricative Plosive – affricate	The boundary is determined by finding the frication section for the junction of the plosive and fricative and then placing the cut just after the release at the start of the frication.	'out~sied ~aak'choo-al ~dout-fal -ad~varns 'aab-sa~luut-ley
Plosive – liquid Plosive – nasal	Like plosive-vowel with the second syllable starting with the onset of vocal vibration immediately following the release of the plosive	~dark-ley -ig~nor-ing ~west-wad 'aab~nor-mal

		~kloud'les ~liet-ning
Fricative – plosive	The boundary is set at the middle of the frication formed by the conjunction of the fricative and plosive. However, when the fricative is an [s] and the plosive is unvoiced then on excision the second syllable sounds as it if starts with the plosive's voiced counterpart. Consequently the second syllable should only be used in this context.	'ko-ris~pond-ing -is~taa-blish ~wenz-day ~fif-tey
Fricative – fricative Fricative – affricate	The boundary is set at the middle of the frication section unless one of the fricatives has a high frequency component, in which case the spectral energy can be used to place the boundary more accurately.	~aat-mas-fia ~kwes-chan-a-bal 'north~see
Fricative – liquid Fricative – nasal	Like fricative-vowel with the second syllable starting with the onset of vocal cord vibration following the release of the fricative.	~harsh-ley ~muuv-ment -un~faiv-ra-bal ~eev-ning ~north-wad ~reez-na-bal
Liquid – plosive Liquid – fricative Liquid – affricate	If the plosive is voiceless then the boundary can be detected as the end of voicing for the liquid. For voiced plosives voicing can occur throughout the boundary region so it may be necessary to look for a sudden rise in amplitude at the end of the liquid to mark the boundary.	~swel-ta-ring -el~swei ~sel-dam -orl~dhoa 'orl-ta~ge-dha ~orl-soa
Liquid – liquid Liquid – nasal	Very few examples of these boundaries in the MeteoSPRUCE database – 5 out of 344. Fixing the boundary for this case is difficult so syllables in these categories are better extracted from other easier contexts and modified as necessary for inserting in this context.	-orl~red-dey ~awl-waiz -fa~mil-ya ~dul-ley ~orl'moast
Nasal – plosive Nasal – fricative Nasal – affricate	Similar to liquid-plosive.	'in-di~kai-shan 'un-da~goa 'aak-si~den-tal 'un~broak-an -kan~sernz ~sen-cha-reez
Nasal – nasal	The nasal region is characterised by a long low amplitude region of the waveform. The middle of this can be taken as the boundary.	'or-tum-nal 'un~noan

Table 1: Rules for identification of syllable boundaries.