AUTOMATIC SYLLABIFICATION USING HIERARCHICAL HIDDEN MARKOV MODELS

Pieter Nel, Johan du Preez

Digital Signal Processing Group
Department of Electrical and Electronic Engineering
University of Stellenbosch, Stellenbosch, South Africa
pieter.nel@za.flextronics.com, dupreez@dsp.sun.ac.za

ABSTRACT

This paper presents a purely statistical method for the automatic syllabification of speech. A hierarchical HMM structure is used to implement a purely acoustical model based on the phonotactic constraints found in the English language. A well-defined DTW distance measure is presented for measuring and reporting syllabification results. We achieve a token error rate of 20.3% with a 42ms average boundary error on a relatively large set of data. This compares well with previous knowledge- and statistically based methods.

1. INTRODUCTION

The syllable was proposed as a unit of automatic speech recognition as early as 1975 due to its strong links with human speech production and perception [1]. It has been suggested that many prosodic properties such as pitch, accent and stress are most naturally expressed in terms of syllables. Some researchers hypothesize the syllable to be the primary unit of segmentation in speech and the basic unit of lexical access in the human mind. It is therefor valuable to be able to automatically syllabify speech. These syllables can be used for TTS, foreign accent identification [2] etc.

Most of the early work on automatic segmentation of speech into syllables used knowledge-based methods. Various algorithms have been proposed to automatically segment speech into syllables of which Mermelstein's convex hull method was one of the first [3]. He achieved a token error rate (TER) of 9.5%, albeit on a limited data set of eleven sentences spoken by only two male talkers. When the same algorithm is used to segment TIMIT into syllables, overall performance dropped to a TER of 26.6% [4]. When modified slightly as reported by Howitt in [4], it improves to a TER of 14.6%.

Recently the focus has been on statiscally-based methods. These have their own inherent problems in that statistical methods are unable to handle conditions that are not present in their training data. Most recently Wu [5] [6] reported a 21% error rate on a subset of OGI Numbers95 using RASTA PLPs as input to a multilayer perceptron.

In section 2 we discuss the syllable definition we conformed to. Next, we describe the speech material in section 3 and in section 4 the recognition system used in our experiments with emphasis on hierarchical HMMs. Finally, in section 5 we discuss the results of our experiments.

Pieter Nel is now with Flextronics SA

2. SYLLABLE DEFINITION

We base our model of the syllable on the commonly accepted perceptual model shown in figure 1 [7].

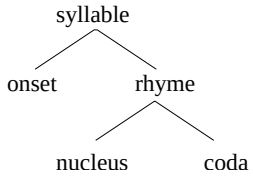

Fig. 1. Syllable parts

The structure in figure 1 when applied to the English language can be represented as $C_0^3 V C_0^3$, where C_0^n signifies 0 to 3 consonants and V signifies a vowel. Employing the phonotactic constraints that apply specifically to English language syllables allow us to further specify it as

$$/s/C_u C_v V C_v C_{u+v+s} C_{u+s}$$

where the members of each group are shown in table 1.

Phonotactic contraints for the English syllable specify that when the onset is 3 consonants long, the first consonant can only be an /s/. According to sonority theory there must be a rising sonority curve in the onset leading up to the nucleus. A further constraint is that, by referring to the list of binary features in table 2, the 2nd consonant must be [-sonorant] and the third [+sonorant][7]. Therefor the /s/ in the onset is followed by the unvoiced consonants C_u and then the voiced consonants C_v.

A syllable must always have at least a nucleus, V, which we define as all vowels, dipthongs and the schwa. Syllabic consonants are treated as /ə/+C.

Group	phonemes
V	/a/ /e/ /i/ /o/ /u/ /aɪ/ /ɛː/ /ɛ/ /ɔː/ /ɔɪ/ /ø/ /ə/ /əː/ /æ/ /œ/ /œː/ /eɪ/ /ui/ /iuː/ /əʊ/ /œy/ /aʊ/ /ɑ/
C_u	/b/ /d/ /f/ /g/ /k/ /p/ /t/ /x/ /θ/ /ts^h/ /ʤ/ /ʃ/ /ʧ^h/ /ʤ/ /ɾ/
C_v	/h/ /j/ /l/ /m/ /n/ /r/ /v/ /w/ /z/ /ð/ /ʒ/ /ŋ/ /R/
S	/s/
C_{u+v+s}	$C_u \bigcup C_v \bigcup S$
C_{u+s}	$C_u \bigcup S$

Table 1. Syllable parts

In the coda we must conform to decreasing sonority. However the sonority generalisation fails to account for one specific class of possible English codas: those with clusters like /sp/ and /sk/ as present in words like *lisp* and *disk*. We therefor include /s/ in the second to last coda position, C_{u+v+s}.

group	description	features
V	vowels	[+*syllabic*] [+*sonorant*]
C_v	voiced consonants	[-*syllabic*] [+*sonorant*]
C_u	unvoiced consonants	[-*syllabic*] [-*sonorant*]
S	/S/	[- *syllabic*] [-*sonorant*]

Table 2. Binary features for syllable classes

Our syllable definition can be applied in defining a regular grammar [8] for the classes in table 1 and 2 as was described by Prinsloo in [9]. This regular grammar has an exact non-deterministic Finite Automoton equivalent which we implement as an HMM.

3. DATABASE

The recognition experiments were performed on a subset of the Sunspeech corpus, a set of continuous, naturally spoken utterances in South African English and Afrikaans. Data was sampled at 16000Hz and recorded in a noise-free environment. It was transcribed by trained linguists on the phone, syllable and word level.

The English subset consists of 40 different sentences spoken by 97 different speakers with a total of 1942 utterances. All sentences are not spoken by all speakers. The data was divided into a training- and test set with 1316 utterances by 66 speakers and 626 utterances by the remaining 31 speakers respectively.

We encountered several inconsistencies in the syllable labels. Single consonants were labelled as syllables, mostly where these consonants are missing the label for a preceding schwa. Many syllables were transcribed containing two vowels. Some examples include:

- single consonants /n/, /v/, /t/, /d/, /ʒ/, /f/, /r/, /k/ labelled as syllables
- words like *reputation* where the last syllable is transcribed as /ʃn/ which does not contain the implicit schwa.
- single /ʧ^h/ /ʦ^h/ /ʤ/
- *about* labelled as single syllable /əbɑt/ therefor containing two nucleii
- *for* labelled as /fr/ therefor missing a nucleus
- *thousands* where the last syllable is labelled as /ʒnʒ/ missing a nucleus
- *evident* split in two syllables where the first is /ɛvə/ again containing two nucleii

These mislabelled syllables were marked by hand in the transcriptions of both our training and test set and ignored in all subsequent experiments.

The five simple syllable structures shown in table 3 account for 94% of all syllables in the Sunspeech database. This is similar to that reported for Switchboard by Wu in [6] where eight relatively simple structures also account for 84% of the syllables found in the corpus.

structure type	% of corpus
V	10.43
VC	13.77
CV	36.25
CVC	28.47
CVCC	5.1

Table 3. Syllable structures in Sunspeech database

4. SIGNAL ANALYSIS AND MODELLING

4.1. Signal processing

After performing preemphasis and energy normalisation in a 100Hz - 7500Hz window, the data was parameterized using 18 dimensional MFCCs with 22 filter banks. A frame length of 20ms with a frame skip of 10ms was used. The delta, and delta of delta between successive frames were computed and the dimensions of the result reduced using linear discriminant analysis (LDA).

4.2. Phones

We trained phone models for 56 distinct phones found in Sunspeech, using both the training and test set. We used a simple left to right HMM structure with one state skip. Since we wanted the best possible input to our syllable model, the phone models were trained using the entire set of training and test data in order to minimise effects due to phone model inaccuracies.

We achieved an accuracy of 53% for all 56 phones tested on the training and test set. Since we build our syllable model using groupings of these phones this level of accuracy was deemed sufficient for our specific set of experiments.

4.3. Hierarchical HMMs

Since the phenomena that we are modelling operates on a number of hierarchical levels, we chose to use a 4 level hierarchical HMM (HHMM) to represent the speech with. The top level represents a speech recording as a combination of syllables and garbage segments as is shown in figure 2. When used to analyse speech, this level of the model generates tags with “syllable” and “garbage” as labels.

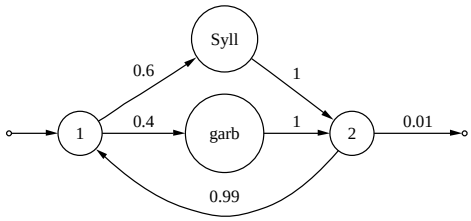

Fig. 2. Segmenter model

The syllable state from figure 2 expands to the second level model as is shown in figure 3. This implements an FSA of the syllable definition described in section 2.

Similarly the garbage state of 2 expands to a 6-state ergodic HMM model on this second level. This garbage model is built using the /S/, C_u, C_v, V classes together with a model for “silence”

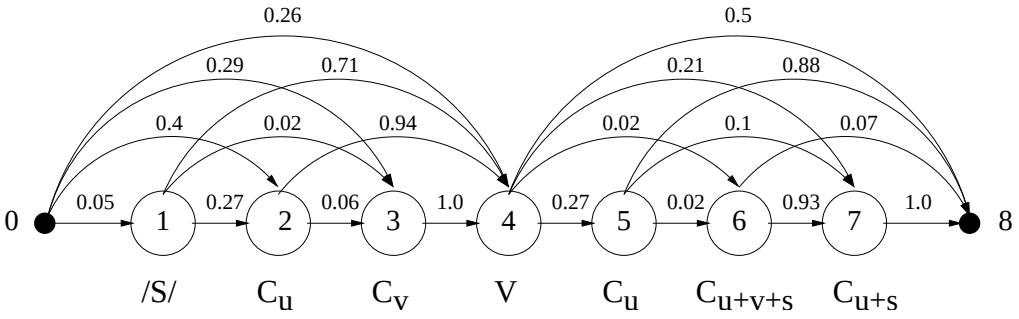

Fig. 3. Syllable model

and one for the "unknown" tag, arranged in a fully connected configuration. (The "unknown" tag found in Sunspeech is a small collection of different phones which was not labelled by the transcribers.)

As shown in figure 4 the third level in the hierarchy models each of the class groups described in table 1. They are built as a parallel combination of their constituent phone models which in their turn form the fourth and bottom-most level in the hierarchy. This level directly interacts with the MFCC feature vectors obtained from the acoustical signal.

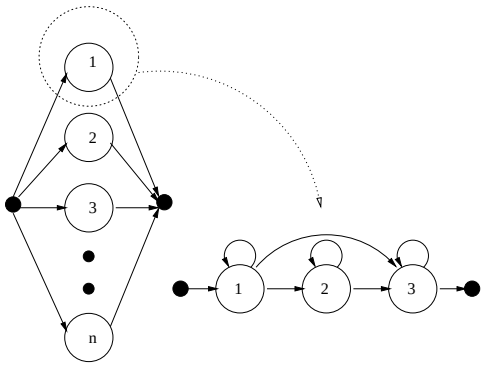

Fig. 4. Parallel HMM model

These phone HMM models are trained separately and then integrated into the HHMM. After this integration their parameters are kept frozen/unchanged with further training impacting only on the higher levels of the HHMM. Specifically the entire syllable HHMM model was trained using the time alligned syllable markings available for the training set. The resulting transition probabilities are shown in figure 3.

5. RESULTS AND DISCUSSION

5.1. DTW distance measure

To allign an automatically determined syllable labelling with its ideal hand-labelled version, a DTW procedure is used to do the mapping between these two sequences in terms of correct labels, substitutions, insertions and deletions[10]. Two components play a role here namely a) the relative costs of these various types of labelling errors, and b) the specific local cost describing how dissimilar a particular label is compared to another. A common problem with automatic syllabification algorithms is that these measures are often described inadequately [4]. We therefor provide the detail of our matching procedure in the following.

We give a small but equal weighting to DTW paths corresponding to substitutions, insertions and deletions (the specific weight was 0.1). Since substitution errors results in a shorter DTW path length than the others, this weighting results in a slight preference for substitution errors compared to insertions and deletions.

Our label distance measure algorithm takes as input the acceptable time error in fixing the boundaries of the syllables. We call this ϵ and used 20ms as our acceptable error margin. Referring to figure 5 we then define the overlap between the original syllable transcription and our generated syllable boundaries.

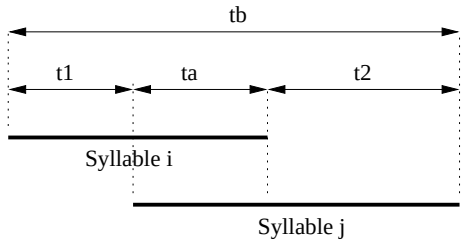

Fig. 5. Syllable overlap

Our distance measure is based on the amount of overlap and whether the generated syllable falls within the accepted boundary error compared to the original transcription.

Definitions:

$$overlap = \frac{t_a}{t_b} \quad (1)$$

$$\epsilon = \text{acceptable boundary error} \quad (2)$$

Step 1:

$$D(i,j) = \begin{cases} 1 - overlap & \text{if } t_1 \text{ and } t_2 < \epsilon, \\ 3 - overlap & \text{if } t_1 \text{ or } t_2 < \epsilon, \\ 5 - overlap & \text{if } overlap > 0 \\ & \text{and } t_1 \text{ and } t_2 > \epsilon, \\ 10 & \text{if } overlap < 0. \end{cases} \quad (3)$$

Step 2:

$$D(i,j) = \begin{cases} D(i,j) + 5 & \text{if ids mismatch} \end{cases} \quad (4)$$

5.2. Syllabification results

Tabel 4 summarizes the results achieved by our automatic syllabification system.

tokens	14143
deletions	12.7%
insertions	5.7%
substitutions	2%
correct	85.4%
accuracy	79.7%
TER	**20.3%**
avg boundary err	42ms
std dev	36ms
max err	406ms

Table 4. Syllabification results

Our token error rate of 20.3% compares well with results obtained by Howitt [4] and Wu [6] on TIMIT and OGI Numbers95 respectively.

We achieved an avarage boundary error of 42ms. With an average English syllable length of 250ms [6] this can be considered fairly accurate.

From the results we have also noticed that, 50% of syllables do not have codas, 13% (0.26*0.5) of syllables only have V, 26% of syllables start with V and 5% of syllables start with /s/. This corresponds well with the characteristics of the hand-labelled version of this database as summarised in table 3

6. FUTURE WORK

We used the Sunspeech database because of its existing hand-labelled syllable-level transcriptions. We intend repeating the experiment on the TIMIT database using Bill Fischer's **tsylb2** program to generate syllable transcriptions. Fischer's program implements the syllable model defined for English by Kahn in [11]. By training our model on this data we will essentially be able to create a statistical representation of Kahn's syllable model as trained on TIMIT. A trained model like this can easily be used as a diagnostic tool to indicate transcription errors when labelling databases.

7. CONCLUSIONS

We have applied the concept of hierarchical HMMs to model syllables. These statistical models are automatically inferred directly from acoustical speech data. It is, however, self-evident that the generelisation ability of these models are highly dependant on the specific training database being used. Evaluation showed the results to be fairly accurate and comparing well to knowledge-based approaches. The ability to observe the resultant regular grammars describing syllable structure also holds some benefit compared to neural-based approaches.

8. REFERENCES

[1] Osamu Fujimura, "Syllable as a unit of speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-23(1), pp. 82–87, February 1975.

[2] Kay Berkling, "Scope, syllable core and periphery evaluation: Automatic syllabification and foreign accent identification," *Speech Communication*, , no. 35, pp. 125–138, 2001.

[3] Paul Mermelstein, "Automatic segmentation of speech into syllabic units," *Journal of the Acoustical Society of America*, vol. 58, no. 4, pp. 880–883, October 1975.

[4] Andrew Wilson Howitt, *Automatic Syllable Detection for Vowel Landmarks*, Ph.D. thesis, Massachusetts Institue of Technology, July 2000.

[5] Su-Lin Wu, Brian E. D. Kinsbury, Nelson Morgan, and Steven Greenberg, "Incorporating information from syllable-length time scales into automatic speech recognition," in *ICASSP*, 1998, vol. 2, pp. 721–724.

[6] Su-Lin Wu, *Incorporating Information From Syllable-length Time Scales into Automatic Speech Recognition*, Ph.D. thesis, University of California, Berkeley, 1998.

[7] Heinz J. Giegerich, *English Phonology - An introduction*, Cambridge University Press, 1992.

[8] Noam Chomsky and Morris Halle, *The Sound Pattern of English*, Harper and Row, Publishers, 1968.

[9] G.J. Prinsloo and M.W. Coetzer, "Automatic syllabification and phoneme class labelling with a phonologically based hidden markov model and adaptive acoustical features," *Computer Speech and Language*, vol. 4, pp. 247–262, 1990.

[10] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon, *Spoken Language Processing*, Prentice Hall, 2001.

[11] Daniel Kahn, *Syllable-based Generalizations in English Phonology*, Ph.D. thesis, Massachusetts Institute of Technology, September 1976.