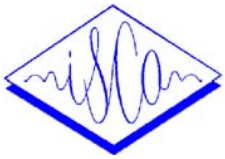

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

HIGH-ACCURACY AUTOMATIC SEGMENTATION

Jan P. H. van Santen[1] *Richard W. Sproat*[2]

[1] Lucent Technologies – Bell Labs, 600 Mountain Ave., Murray Hill, NJ 07974, U.S.A., jphvs@research.bell-labs.com
[2] AT&T Labs – Research, 180 Park Ave., Florham Park, NJ 07932-0971, U.S.A., rws@research.att.com

10.21437/Eurospeech.1999-620

ABSTRACT

We propose a system for automatically determining boundaries between phonetic segments in a speech wave given a phonetic transcription: automatic segmentation. The system uses edge detectors that are applied to various speech representations; both are optimized for each diphone or diphone class. Output from these detectors, which contains spuriously detected edges, is then combined with alternative pronunciations generated via rules from the canonical pronunciation. The final output is generated with lowest-cost path algorithms applied to finite state transducers.

1. INTRODUCTION

Automatic segmentation is critical both for speech research and for speech technologies that rely on segmented speech corpora for training or construction purposes. For example, in text-to-speech synthesis (TTS) segmented corpora are used for the construction of intonation, duration, and synthesis components [4].

The standard approach to automated segmentation is to adapt an automatic speech recognition (ASR) system by restricting its language model to the known input sentence [2, 6]. Results from these methods are quite impressive, but there are also serious shortcomings that prevent them from reaching the important goal of being able to fully automatically generate high-quality TTS from unannotated speech corpora.

First, many TTS system require boundaries to obey standard phonetic conventions. Unless ASR systems are trained on segmented (not only labeled) speech and unless proximity to the boundaries is part of the optimality criterion, these systems may put boundaries at quite different locations such as, for vowel - voiced fricative boundaries, the onset of frication instead of formant structure. These conventions are important, e.g. because the durational behavior [1, 5] of conventionally segmented speech is better understood than that of speech segmented differently. Also for concatenation one needs to rely on algorithms that presume a certain internal acoustic structure of acoustic units (e.g., a voiceless stop initial unit should start with a silence).

Second, the accuracy levels required for high-quality TTS applications must be quite high. For example, if a system has an average error of 10 ms in detection of voiced closure - burst boundaries, then this may result in voiced stop - vowel acoustic units that miss the burst altogether because these bursts often last less than 10 ms.

A third shortcoming is that ASR based systems require large amounts of language-specific training data. In the case of multilingual TTS, such data are often not available. Creating these data would defeat the purpose of saving manual labor.

In this paper, we propose a system that attempts to circumvent these shortcomings. We surmise that the deeper reason that ASR-based systems may not necessarily provide the best route towards high-accuracy segmentation is that they are built to *identify* phonetic segments, not to *detect boundaries between* phonetic segments. Not surprisingly, our approach is complementary to ASR-based approaches in that it focuses on boundary detection (using edge detection techniques borrowed from image processing) while at the same time grouping certain diphones together in classes sharing the same acoustic model.

2. PROPOSED SYSTEM

2.1. Broad and Narrow Diphone Classes

The acoustic manifestation of a given phonetic segment is influenced by its phonetic context. E.g., a /b/-closure tends to be voiced when preceded by a voiced sound (as in "hezbollah"), and voiceless otherwise (as in "softball"). As a result, boundary detection using independently constructed z-end (or t-end) and b-start detectors is unlikely to work well. In addition, which acoustic features are critical depends on both phones flanking the boundary. For example, an f-s boundary is best detected by looking for decreased energy below 2,000 Hz and increased energy above 4,000 Hz, while an f-**V** boundary is best detected by measuring changes in the 800 - 2,500 Hz band. Thus, accurate localization of a boundary between two phonetic segments depends on the acoustic characteristics of both segments and on what the key contrast between them is. The number of possible diphones is more than 1,000 in most languages, which poses a threat to our goal to minimize the amount of training required. We have observed, however, that the set of all diphones can be partitioned in two types

of *diphone classes*: *broad* and *narrow*.

Broad diphone classes are based on the standard categorization of phonetic segments in terms of manner and production and voicing. In American English, we distinguish between the following natural classes: **B**: *Voiced stops* (closure vs. bursts), **J**: *Voiced affricates* (closures vs. fricative), **Z**: *Voiced fricatives*, **P**: *Voiceless stops* (closure vs. bursts), **C**: *Voiceless affricates* (closures vs. fricative), **S**: *Voiceless fricatives*, **N**: *Nasals*, **L**: *Liquids*, **H**: *h*, **G**: *Glides*, **V**: *Vowels*, **v**: *schwa*. We use the term *broad diphone class* to denote a set of diphones whose respective members belong to the same respective classes. For example, **S-V** is a broad diphone class, with as typical members *s-i* and *f-u*.

A key feature of most boundaries between phonetic segments belonging to different natural classes is that they can be characterized by *energy changes in broad regions of the spectrum*. For example, boundaries between a voiceless fricative and a vowel involve a sharp rise in energy below 2,500 Hz and a decrease (followed by an increase for stressed vowels) above 4,000 Hz, regardless of which individual fricative or vowel is involved.

The remaining diphones must be treated on an individual basis. Boundaries between specific glides and vowels primarily involve formant movement (but a different movement depending on the individual segments involved), boundaries between nasals involve broad band energy changes but in bands that differ for different nasal - nasal diphones), and boundaries between vowels are characterized by either specific formant movement, insertion of a glide, or insertion of a glottal stop, where the latter can range from a complete silence to a barely measurable attenuation. In all these cases, we have to use different boundary detectors for different diphones. We call these singleton classes *narrow diphone classes*.

2.2. Multiple speech representations

The distinction between narrow and wide diphone classes suggests using different speech representations for different diphone classes. Specifically, it suggests that perhaps energy in broad bands should be used for broad diphone classes, formants or cepstrum parameters for vowel-vowel boundaries, and yet other representations for *n-n* or *s-f* boundaries. In our current implementation, we use only two representations: 5-band (*5B*) and 55-point mel FFT (*55MFFT*) representations.

For the 5B representation, we process speech sampled at 12KHz with band-pass filters having edge frequencies of B_0 = (100,300), B_1 = (300,800), B_2 = (800,2500), B_3 = (2500,3500), and B_4 = (3500,5800). The subscripts roughly refer to corresponding subscripts for formants. The signals in the five bands are squared, smoothed using 1 ms Hamming windows at 1 ms intervals, and transformed into dB via 20 log_{10}. For the 55MFFT representation, we used FFT sampled at 55 points equally spaced on the mel scale.

2.3. Broad band edge detection

We observed that energy changes tend to have one of two shapes: *sigmoid*, with a sharply localized area of steepest ascent or decent; or *folded*, with a sudden change in direction but without a well-defined area of steepest descent. Examples of the former can be found in bands B_0 through B_3 of the *s-e* transition, and an example of the latter is in B_4 of the same transition where a near-linear descent during the final 20 ms of the /s/ is followed by a sudden leveling out (for unstressed /e/) or near-linear increase (for stressed /e/). The point where the direction change takes place is well defined, but there is no clear steepest point.

We detect both types of edges by convolving the speech signal in each band with an appropriately chosen wavelet. We use symmetric (for folded edges) and antisymmetric (for sigmoid edges) Gabor functions:

$$(k/\sigma)^2 e^{-0.5(kt/\sigma)^2} cos(kt) \text{ and}$$
$$(k/\sigma)^2 e^{-0.5(kt/\sigma)^2} sin(kt)$$

With properly selected values for k and σ, the result of convolution tends to show sharp peaks at edge locations. We subtract constants from the symmetric Gabors so that the sum of their values are 0 (this is necessarily the case for antisymmetric Gabors). The zero-sum property in conjunction with the dB transformation ensures that detector output is *DC free*.

Asynchronous edge detection For a given transition class, we specify for each band which edge detector (symmetric, antisymmetric, k positive vs. negative, and value of σ) is to be used, and how this band should be weighted in the final decision. There are many ways in which detector outputs from different bands can combined. One simple way would be to add the outputs for diphone class d, $Out_0(t;d), \cdots, Out_4(t;d)$ and compute the peak location of:

$$Out(t;d) = \sum_{i=0}^{i=4} Out_i(t;d) \qquad (1)$$

But this creates an interesting problem. Because many boundaries involve actions of multiple, imperfectly coordinated articulatory events, edges do not occur at the same time across frequency bands and in fact may show different locational constellations in different instances of the same diphone. We have seen instances of the *s-e* boundary where a sharp increase

in B_0 preceded the direction change in B_4 (see above) by 10 ms, and instances where they coincided. This has to do with the weak coupling of the events of voicing onset, termination of frication due to the tongue being withdrawn from the alveolar ridge, and start of frication due to breathiness (in stressed vowels). For this reason, we compute

$$Out(t;d) = \sum_{i=0}^{i=4} \sum_{\tau=t-N}^{\tau=t+N} W(\tau) Out_i(\tau;d) \qquad (2)$$

where $W()$ is a Gaussian window. This has the effect of producing a higher peak the more closely spaced and the larger the per-band peaks are; but it does not require exact simultaneity of these peaks, as in Eq. (1).

2.4. Narrow band edge detection

We compute the vector cross product $\lambda_d^t \mathbf{B}_{55}(t)$, where $\mathbf{B}_{55}$ is the 55MFFT signal, and where λ_d is a vector of weights characteristic of diphone d. The vector λ_d defines a direction in 55-dimensional space on which training samples of the two phones in d, d_1 and d_2, are maximally different (i.e., the first linear discriminant axis). The procedure attempts to find the time point at which, along this direction, the speech signal suddenly becomes much more similar to d_2 than to d_1. Calling the centroids of these two phones $\mathbf{B}_{55}(d_1)$ and $\mathbf{B}_{55}(d_2)$, and their projections on the discriminant axis $x_1 = \lambda_d^t \mathbf{B}_{55}(d_1)$ and $x_2 = \lambda_d^t \mathbf{B}_{55}(d_2)$, we compute the absolute differences $\delta_1(t)$ and $\delta_2(t)$ between $\lambda_d^t \mathbf{B}_{55}(t)$ and x_1 and x_2. We then consider the curve

$$C(t) = (\delta_1(t) - \delta_2(t))/(\delta_1(t) + \delta_2(t)) \qquad (3)$$

We found that $C(t)$ generally has a sigmoid shape in the proximity of the boundary, whose steepest point can be detected with the antisymmetric wavelet. Earlier attempts with the simpler procedure which attempts to find the point where $\delta_1(t) = \delta_2(t)$ failed because such points often do not exist.

2.5. Recognition information

Despite isotonic smoothing, the detector for a given diphone or diphone class is likely to produce spurious peaks, although there is a strong tendency for the largest peaks to occur at the correct locations. We counter these spurious peaks by applying a secondary acoustic model, which simply consists of the vector:

$$B(t) = (B_0(t-10), \cdots, B_4(t-10), B_0(t+10), \cdots, B_4(t+10))^t \qquad (4)$$

We subtract the mean value from B, again to ensure DC-independence, and discard the tenth component to avoid singularity of the covariance matrix (see below), resulting in vector $b(t)$. At time t, we compute the multivariate normal density $G_d()$ for $b(t)$ for diphone class d based on centroid μ_d and full covariance matrix Σ_d. In other words, we borrow here some of the standard steps used in ASR, but using a very coarse speech representation (5 broad frequency bands as opposed to dozens of cepstrum, delta-cepstrum, and delta-delta cepstrum parameters).

Using Bayes rule, and after proper normalization, we compute the *overall acoustic cost* $C(t;d)$ for diphone class d at time t based on this density and on peak height, $Out(t;d)$.

Thus, the end product of the acoustic analysis is a cost function $C(t;d)$ at each time point t and for each diphone class d, which reflects both the strength of the output of the detector for diphone class d and the degree to which the spectrum locally resembles the typical spectrum for d.

2.6. Multiple pronunciation modeling via rewrite rules

Small deviations from the canonical pronunciation can have large effects on system performance. For example, if a voiced sound is pronounced without voicing (e.g., the /z/ in the word "using" is pronounced as /s/), the system may look in vain for transitions of the wrong type.

The solution we have opted for is to use explicit *alternative pronunciation rules* of the type

$$z \rightarrow s < 2.0 > | z < 1.0 > / V\_V.$$

I.e., "/z/ becomes /s/ with cost 2.0 and remains a /z/ with cost 1.0, in intervocalic context". We apply these rules to the canonical pronunciation to generate a lattice of legal alternative pronunciations.

A fundamental limitation of this approach is that in complicated sentences the combinatorics becomes unwieldy. Yet, the rules may still not cover the range of pronunciations that may occur. An additional problem is the estimation of these pronunciation costs.

2.7. Joint optimization of pronunciation and acoustic costs

Both the acoustic cost table and the lattice of alternative pronunciations generated by the rules from the canonical pronunciation can be compiled into finite state transducers, and compiled into a single transducer for which we can then compute the lowest cost path. This path lists a sequence of phone boundaries that corresponds to one of the legal alternative pronunciations and has the lowest overall (acoustic + pronunciation) cost.

3. SYSTEM TRAINING

3.1. Broad band detectors

We used a hand segmented single speaker speech corpus of about 2,000 sentences. For a given broad band detector for diphone class d, we applied all 64 combinations of 16 sizes, two shapes (antisymmetric vs. symmetric Gabor), and two signs (+/-) to each band of all instances of the d diphone class, and measured the root mean squared deviation of peak locations from the manual boundary. We selected for each band the optimal combination, and computed a weight that was inversely related to the deviation.

While coarse, this training method is extremely fast, and is more robust than several other methods we have experimented with. Also note that the total number of parameters per diphone class is quite small.

3.2. Narrow band detectors

For a given diphone $d_1 - d_2$, we collect two sets of 55MFFT vectors: from the final 25% time interval of occurrences of d_1 where d_1 is followed by segments having similar formant characteristics as d_2, and from the initial % time interval of occurrences of d_2 where d_2 is preceded by segments having similar formant characteristics as d_1. Thus, for the $y - o$ diphone we may consider y followed by u or o, and o preceded by i or y. We next perform a linear discriminant analysis on the pairs of vectors to produce the vectors λ_d.

4. SYSTEM PERFORMANCE

On the training data, we found superior performance on many classes, including *B-V*, *Bcl-B*, *N-V*, *B-L*, *P-S*, *S-V*, with 95% of errors being less than 6 ms, and 50% less than 2 ms. Informal observations using a new speaker confirmed these performance levels. For the traditionally difficult cases such as *G-V*, we found 50% of errors at less than 5 ms, but at least 10% of errors in excess of 10 ms. Also these results were confirmed.

5. CONCLUSIONS

Although the current results are quite encouraging, and reach levels comparable to the best human segmenters, these results are obviously tentative because they are based on single-speaker training data. We are currently evaluating the system for other speakers and also for other languages (Hindi), and will report results in the near future.

We view this system as a first step towards systems that have as their central building block a detector that is specialized for a specific diphone or diphone class: for different classes, different speech representations may be used, and the detection mechanism (e.g., edge detection vs. linear discriminant analysis based zero crossings) may also differ. By using phonetic categories that are based on basic articulatory notions (see, e.g., [3]), are well understood and, with some effort, can be mapped on different languages, we believe it is possible to build a system that can be adapted to new languages with little new training data.

In our experience, the main reason for system failure is the inability of the alternative pronunciation rules to capture the full range of speaker variability. Among the many details that need work are stress-specific detectors, exploring other representations (e.g., cepstrum instead of 55MFFT), using ASR as pre-processor, or as replacement for our 5-band based recognition process in Section), and more extensive training on multiple speakers to fill in the gaps the current system has due to data sparsity.

6. ACKNOWLEDGMENTS

We thank Oded Ghitza, Bernd Möbius, Bhuvana Narasimhan, Joseph Olive, and Chilin Shih for helpful discussions. Also, finite state machine software written by Michael Riley was critical for the success of this project.

7. REFERENCES

1. J. Allen, S. Hunnicutt, and D. Klatt. *From text to speech: The MITalk system.* Cambridge University Press, Cambridge, 1987.
2. A. Ljolje and M. D. Riley. Automatic segmentation of speech for TTS. In *Proc. of Eurospeech-93*, volume 2, pages 1445–1448, Berlin, 1993.
3. J. Olive, A. Greenwood, and J. Coleman. *Acoustics of American English speech: a dynamic approach.* Springer-Verlag, New York, 1993.
4. R. W. Sproat. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach.* Kluwer, Boston, MA, 1997.
5. J. van Santen. Timing. In R. Sproat, editor, *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, chapter 5, pages 115–139. Kluwer, Boston, MA, 1997.
6. C. W. Wightman and D. T. Talkin. The Aligner: Text-to-speech alignment using markov models. In J. van Santen, R. Sproat, J. Olive, and J. Hirschberg, editors, *Progress in Speech Synthesis*, pages 313–324. Springer-Verlag, New York, 1996.