Acoustic cues identifying phonetic transitions for speech segmentation

D.R. van Niekerk and E. Barnard

Human Language Technologies Research Group, Meraka Institute, CSIR, Pretoria /
School of Electrical, Electronic and Computer Engineering,
North-West University, Potchefstroom, South Africa
dvniekerk@csir.co.za, ebarnard@csir.co.za

Abstract

The quality of corpus-based text-to-speech (TTS) systems depends strongly on the consistency of boundary placements during phonetic alignments. Expert human transcribers use visually represented acoustic cues in order to consistently place boundaries at phonetic transitions according to a set of conventions. We present some features commonly (and informally) used as aid when performing manual segmentation and investigate the feasibility of automatically extracting and utilising these features to identify phonetic transitions. We show that a number of features can be used to reliably detect various classes of phonetic transitions.

1. Introduction

Defining exact boundaries between phonetic segments in speech is difficult, especially in those contexts where co-articulation between neighbouring phones renders boundary definition somewhat ambiguous. Nevertheless, for the purposes of spoken language research and system development, a pragmatic approach is necessary in order to define such boundaries as accurately and consistently as possible. Research into the development of corpus-based text-to-speech systems has suggested that consistency (in addition to accuracy) of boundary placements is an important factor when considering the eventual quality of these systems [1, 2].

Most early development of speech corpora involved manual effort by language or phonetics experts with a significant amount of experience in identifying phonetic segments from visual and auditory information. This reliance on expert human involvement has endured, despite advances in speech recognition and machine learning techniques applied to automating this task. As much is evident when one considers that high quality corpora are still manually checked by such individuals [3].

The expert manual transcription procedure can be viewed as a two-stage process, where the transcriber initially identifies segments based on the acoustic properties (aided by visual representations thereof) and subsequently refines boundary placements between contiguous segments by considering sets of consistent acoustic cues based on the transition context (defined by broad phonetic classes).

The application of Hidden Markov Models (HMMs) to phonetic segmentation can be likened to the first stage of the expert procedure described above and in cases where such models are sufficiently trained, this leads to boundary placements which for the most part are fairly similar to the "ideal" locations [4]. This is especially true when manually segmented data exists with which to bootstrap the process involved in training HMMs. Nevertheless, a large amount of research has been done on further reducing the discrepancies between HMM based and manually obtained boundaries (i.e. "boundary refinement") [5, 6, 7]. This has been justified by the observation that manually segmented and refined automated methods usually result in better quality synthesis when compared to baseline methods [8, 9]. The implementation of the boundary refinement stage has largely involved the application of statistical machine learning techniques relying on samples of manually segmented data in order to "learn" the conventions of expert transcribers without explicitly considering the underlying process or considerations taken into account. This has proved successful, with researchers reaching levels of accuracy rivalling what can be expected when compared to discrepancies between independently verified alignments by experts [5].

Unfortunately, the feasibility of applying techniques such as these is limited in the context of developing corpora toward building systems for languages where resources and expertise are scarce. This is the case for two primary reasons:

- Corpora are designed minimally in order to minimise effort in text selection (it is difficult to find reliable electronic texts for these languages) and expertise required during recording and annotation. This results in corpora where some phonetic contexts simply do not have sufficient observations in order to train adequate acoustic models.
- No manually checked corpora pre-exist in most of the languages of the developing world, because of a lack of skilled persons to perform such tasks. Corpora which are hand checked are small and have mostly been produced by persons with limited background and training.

For the purposes of developing relevantly annotated corpora with the goal of building high quality spoken language systems, it is thus worthwhile investigating the automated extraction and application of acoustic cues to identify phonetic transitions in much the same way as a human transcriber would. To this end we identify important features and the feasibility of extracting phonetic events from such features. The identification of reliable acoustic cues would have the following advantages for automated corpus development:

- Boundary candidates obtained in this way can serve as an independent point of reference for judging the quality of alignments (whether automatically or manually obtained).
- These boundary candidates can be integrated into an automated procedure in order to refine boundary placements or improve the quality of training acoustic mod-

els, taking into consideration a specific protocol with the end goal of the segmented corpus in mind.

In this paper we present an initial analysis of the effectiveness of various cues for detecting phonetic events in different contexts in order to determine the feasibility and potential impact of applying this information. Section 2 describes the identification of potential features, Section 3 describes the experimental setup including the details of identifying boundary candidates. Finally, we report on the results obtained (Section 4) and conclude with a discussion in Section 5.

2. Acoustic features

In large resource collection efforts the development of annotated corpora has typically been realised by the collaboration of a large number of trained individuals. The collaboration of multiple individuals is essential in order to complete the sizable task of manually verifying the quality of phonetic alignments within acceptable time-frames, and to have reliable methods of quality assurance.

Due to the ambiguities which exist at phonetic transitions, it is common to define protocols for the placement of phonetic boundaries based on broad phonetic class categories in order to ensure the consistency of the end result across different individuals [10, 3].

Typical protocols incorporate practical guidelines for the identification of phonetic boundaries based on acoustic cues exhibited by various features that can be extracted or calculated and displayed. This includes the signal energy, estimated fundamental frequency, periodicity (voicing), extracted formant contours, spectral characteristics and waveform shape. Instructions on boundary placement range from complex and highly conditional (e.g. when transcribing approximants, some suggest observing the formants, F3 and F4 for "energy reduction") to relatively simple and clearly defined (e.g. place a phonetic boundary "just prior to the burst of energy" when transcribing a stop consonant). Considering this and initial experiments on how reliably one can estimate or extract all of these features, we have concentrated on the following features for the automatic identification of segmentation cues:

- Signal intensity,
- Fundamental frequency (f0),
- Signal envelope, and
- Cepstral distances.

Due to difficulties in reliably determining the number of formants present as well as the exact contours, we have chosen to rely on the use of a "cepstral distance" measure (defined in Section 3.3.5) which we hope will identify changes in the formants and general spectral changes with sufficient accuracy.

3. Experimental setup

We employed the *Praat* [11] and *HTK* [12] software packages to aid in extracting features from three sets of manually annotated audio recordings representing typical minimally designed TTS corpora (see Table 1).

Language	Gender	Utterances	Duration	Phones
Afrikaans	Male	134	21 mins.	12341
isiZulu	Male	150	19 mins.	8559
Setswana	Female	332	44 mins.	26010

Table 1: Reference data sets

3.1. Broad phonetic classes

The most practical and relevant view of phonetic transition contexts for this study is based on broad phonetic categories. All segment labels in the above-mentioned corpora are thus mapped to one of the following labels in accordance with International Phonetic Alphabet (IPA) definitions: *affricate*, *approximant*, *click*, *fricative*, *nasal*, *pause*, *stop*, *trill* and *vowel*.

The *pause* label is used both with reference to long pauses (typically only occurring at the beginning and end of utterances) and short segments associated with little signal energy such as glottal stops and closures.

3.2. Generating boundary candidates

In general, boundary candidates are established by firstly calculating or estimating contours for the particular feature and either using this contour directly where applicable or deriving a subsequent contour representing the slope by means of numerical differentiation. After obtaining the appropriate representation, we employ a simple peak detection algorithm in order to generate boundary candidates at specific time instants. We briefly present these methods below.

3.2.1. Numerical differentiation

In order to obtain a relatively smooth contour suitable for subsequent peak detection to be effective, we firstly calculate the difference between each sample of the original contour x to obtain a new sequence of differences x_d defined for time instants in-between the original time instants. An odd number N of "difference samples" are framed resulting in a frame x_{df} for each time instant. From this the gradient is determined by first windowing the frame with a simple exponential window function:

$$w[n] = 2^{-|n-\frac{N-1}{2}|}, \qquad (1)$$

obtaining a frame with weighted differences x_{df_w}:

$$x_{df_w}[n] = x_{df}[n]w[n], \qquad (2)$$

and calculating the slope at t (the time instant at the center of the frame) by averaging the weighted differences in each frame:

$$x'[t] = \frac{1}{N}\sum_{n=1}^{N} x_{df_w}[n]. \qquad (3)$$

3.2.2. Peak detection

For detecting local extrema that are of interest during candidate identification, we frame the relevant contour, obtaining an odd number of samples that constitute each frame and simply flag the time instant of the central sample within the frame if it is a global extremum within the frame.

3.3. Acoustic cues

Taking into account the observations in Section 2 we experimented with extracting features and identifying candidates automatically. We now briefly describe the particular cues investigated.

3.3.1. Intensity dynamics

It was observed that many phonetic transitions coincide with changes in the signal intensity and initial experiments indicated

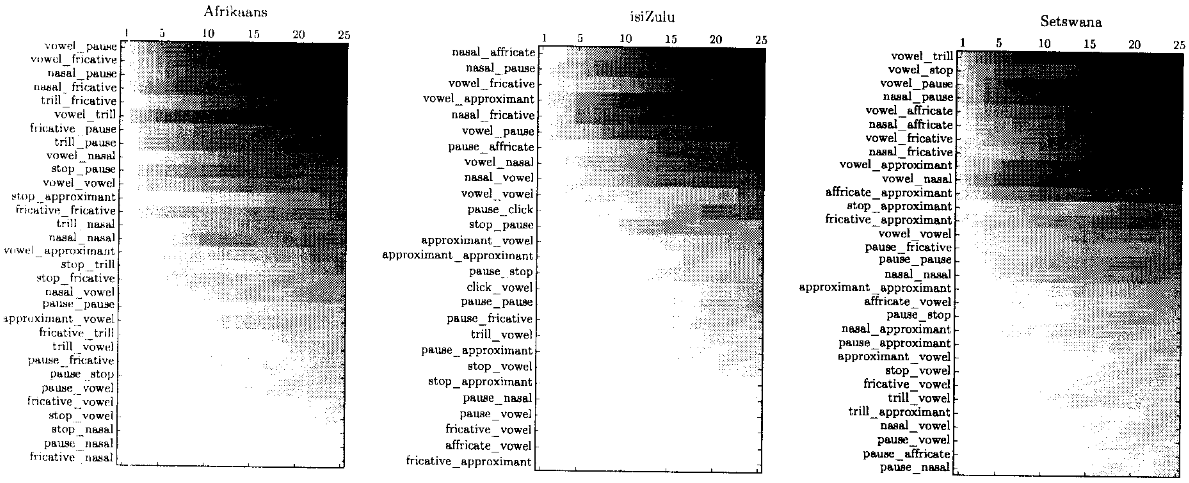

Figure 1: Detection rates: for each phonetic transition context we obtain detection rates for a range of time thresholds (in milliseconds), darker areas represent higher detection rates; this figure represents rates when using the intensity gradient minima cue for each of the languages.

that the slope of the intensity contour peaked near potential boundaries. We thus determine intensity values at 5ms intervals and subsequently obtain the derivative and flag the local minima and maxima of the resulting contour (we distinguish between candidates at minima and maxima).

3.3.2. Waveform envelope

Between neighbouring voiced regions such as vowels and nasals, "dips" in the waveform can indicate a phonetic transition. By obtaining the waveform envelope and flagging local minima, such events can be detected. The use of the intensity contour directly was considered, but in cases such as just mentioned, the envelope provides a more pronounced cue.

3.3.3. Voicing

By means of a pitch analysis in the frequency range 75Hz to 600Hz, one obtains regions that have a strong periodic component which can be identified as voiced regions. By distinguishing between periodic and aperiodic regions one can place boundary candidates between neighbouring regions in the hope of detecting transitions between voiced and unvoiced segments.

3.3.4. Fundamental frequency dynamics

It has been noted that there exists structure within the f0 contour which can be used to identify phonemic events [9]. We attempt to detect these events by employing the *Praat* pitch detection algorithm [13] in the 75Hz to 600Hz range and analysing the slope of the resulting contour.

3.3.5. Cepstral distance

As a measure of spectral difference, which is often used directly via observing the spectrogram or more specifically the changes in formants in order to identify boundary locations manually, we calculated 12 mel frequency cepstral coefficients in 20ms windows with a 2ms time shift. Using this observation sequence we consider windows of N observations, calculate the average of the first $N - 1$ observations and simply calculate the euclidian distance between the last observation and the average calculated in order to obtain a contour representing a measure of difference between each observation and the prior $N - 1$ observations. This contour exhibits peaks at points where the spectral properties change radically.

3.4. Evaluation metric

Because boundary candidates will not coincide exactly with reference boundary locations, we consider a reference boundary location to be *detected* when a candidate boundary is located within a certain time threshold of the reference (following a strategy similarly defined in [14]). Subsequently we define an *unambiguous detection* where only detections with at most one candidate within the defined window around the reference are considered. This discredits detections where false alarms are present. For a specific phonetic transition context we can thus define the *unambiguous detection rate* as the ratio between the number of unambiguous detections and the number of occurrences for each context.

4. Results

By analysing the detection rates for various cues and phonetic contexts over a range of time thresholds, it is possible to obtain a detailed picture of the success of each cue based on phonetic context (see Figure 1 for an example). To investigate the detection rates for individual phonetic contexts, we have to evaluate a range of time thresholds instead of one common threshold (such as 20ms, which is often used), because of the relative durations of phones (e.g. stop phones often have average durations of less than 20ms).

In the subsequent sections we present quantitative results obtained when applying the techniques described on the corpora mentioned in Section 3 (see Table 1).

4.1. Transition detection: coverage

To measure the utility of each cue, the number of detections as a percentage of the total number of transitions is determined. This is done by firstly distinguishing contexts which are deemed successfully detected in general (it was decided that any transition context with detection rates in excess of 70% would be considered), after which detections are summed for these contexts. The results of this process are presented in Table 2.

Cue	Afrikaans	isiZulu	Setswana
Intensity gradient maxima	39.8%	49.1%	38.1%
Intensity gradient minima	36.4%	28.9%	37.4%
Cepstral difference	32.3%	53.5%	35.2%
Waveform envelope minima	36.9%	33.0%	52.8%
Voicing	4.4%	5.8%	37.5%
F0 gradient extrema	3.6%	10.0%	17.9%

Table 2: Cue significance: the percentages reflect the fraction of all phonetic transitions which are successfully detected by each of the listed cues; only transition contexts for which at least 70% detection is achieved are included in these counts.

By using the same notion of successfully detected context, it is also interesting to note the combined transition coverage by the complete set of cues. Figure 2 shows the cumulative coverage when the total occurrences for successfully detected phonetic contexts by each cue are added in turn.

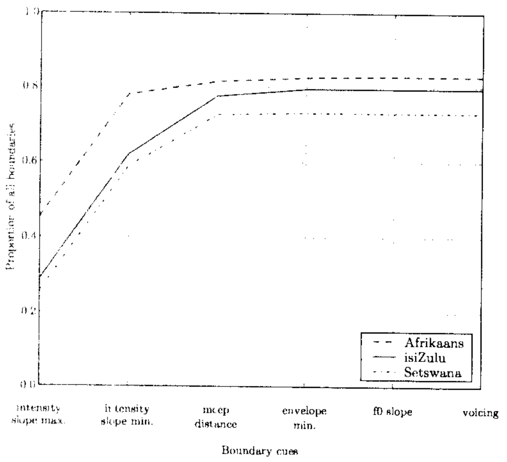

Figure 2: Coverage: the graphs represent the fraction of all phonetic transitions when the number of occurences of successfully detected transition contexts are accumulated for each language.

4.2. Problematic contexts

By differencing the set of contexts that are successfully detected with the complete set, the set of contexts which are least successfully detected is obtained (listed in Table 3). The sets obtained are not surprising considering most of the contexts listed are generally found to be relatively ambiguous (e.g. approximant-vowel transitions) and difficult to distinguish even by manual transcribers. Some of the contexts listed here are also relatively short in duration which suggests that the candidate generation methods used might not be well suited to these conditions.

Afrikaans:
stop-fricative, stop-trill, stop-pause, vowel-nasal, trill-approximant, fricative-pause, approximant-vowel, trill-stop, nasal-nasal, vowel-approximant, fricative-fricative
isiZulu:
pause-affricate, stop-approximant, approximant-pause, affricate-pause, approximant-vowel, stop-vowel, vowel-vowel
Setswana:
pause-affricate, stop-approximant, trill-pause, trill-approximant, approximant-pause, nasal-trill, trill-trill, stop-stop, affricate-pause, approximant-vowel, affricate-affricate, stop-vowel, fricative-nasal, vowel-vowel, pause-trill, fricative-fricative

Table 3: Problematic transition contexts: the contexts listed here were not successfully detected by any of the cues investigated.

5. Conclusion

In this paper we demonstrated the possibility of generating phonetic boundary candidates based on specific acoustic cues that were extracted for three different languages. We showed that it is possible to detect actual boundary positions to a large degree (especially in contexts where the specific cue is relevant from the perspective of speech production).

Although each cue had specific contexts where it outperformed others, the most significant cues were based on the intensity contour and cepstral distance. The fundamental frequency proved to be less successful than expected (based on [9]), but this can probably be attributed to the nature of the reference TTS corpora where the tone is kept more constant than in purely natural speech. Another interesting observation is that the voicing cue worked reasonably well for the female voice but poorly for the male voices, based on these results one should probably carefully consider the exact pitch range of the specific voice before attempting to use this cue.

The problematic contexts remaining seem to be either acoustically ambiguous (e.g. approximant-vowel boundaries cannot be easily distinguished by spectral properties or by observing the waveform) or present cases where our method of candidate generation fails. Segments with very short durations can cause the peak detection method or averaging process set up for the average case to miss detections and particularly the cepstral distance measure proposed would also be more effective for longer segments. Future work in detecting the remaining transitions might involve more sophisticated candidate generation or the application of more appropriate features (formant contours might prove successful).

The identification of boundary candidates presented here will allow us to improve the quality of the alignment process automatically. This can be done by defining a protocol similar to protocols designed to allow consistency between multiple human transcribers and using this directly or integrating candidates into training procedures in order to refine models with respect to precise boundary placements. Another useful application would be to flag potentially misaligned boundaries during quality control of manually or automatically segmented corpora.

An important observation is that boundary refinement based on these candidates can be done automatically and with the target use in mind. This presents opportunity for further research questions relating to text-to-speech synthesis quality when relying on certain acoustic cues to define boundaries. Important acoustic properties relating to speech parametrisation used for speech synthesis should also be explored, e.g. when employing the Harmonics Plus Noise Model, the maximum voiced frequency contour might prove relevant when performing segmentation.

6. References

[1] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[2] M. Makashay, C. Wightman, A. Syrdal, and A. Conkie, "Perceptual evaluation of automatic segmentation in text-to-speech synthesis," in *Proceedings of ICSLP*, Beijing, China, October 2000, vol. 2, pp. 431–434.

[3] M. A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, "The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability," *Speech Communication*, vol. 45, no. 1, pp. 89–95, 2005.

[4] D.R. van Niekerk and E. Barnard, "Important factors in HMM-based phonetic segmentation," in *Proceedings of PRASA*, Pietermaritzburg, South Africa, November 2007, pp. 25–30

[5] D.T. Toledano, L.A. Hernández Gómez, and L.V. Grande, "Automatic Phonetic Segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 617–625, 2003

[6] A. Sethy and S. Narayanan, "Refined Speech Segmentation for Concatenative Speech Synthesis," in *Proceedings of ICSLP*, Denver, Colorado, USA, September 2002, pp. 149–152.

[7] Y. Kim and A. Conkie, "Automatic Segmentation Combining an HMM-Based Approach and Spectral Boundary Correction," in *Proceedings of ICSLP*, Denver, Colorado, USA, September 2002, pp. 145–148.

[8] J. Adell, A. Bonafonte, L.A. Hernndez Gmez, and M. J. Castro, "Comparative study of automatic phone segmentation methods for tts," in *Proceedings of ICASSP*, Philadelphia, Pennsylvania, USA, Mar. 2005, vol. 1, pp. 309–312.

[9] T. Saito, "On the use of F0 features in automatic segmentation for speech synthesis," in *Proceedings of ICSLP*, Sydney, Australia, December 1998, vol. 7, pp. 2839–2842.

[10] R. Cole, M. Noel, and V. Noel, "The CSLU Speaker Recognition Corpus," in *Proceedings of ICSLP*, Sydney, Australia, December 1998, pp. 3167–3170.

[11] P. Boersma, *Praat, a system for doing phonetics by computer*, Amsterdam: Glott International, 2001.

[12] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Veltchev, and P. Woodland, *The HTK Book (for HTK Version 3.3)*, Cambridge University Engineering Department, http://htk.eng.cam.ac.uk/, 2005.

[13] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Insitute of Phonetic Sciences, University of Amsterdam*, Amsterdam, The Netherlands, 1993, vol. 17, pp. 97–110.

[14] Y. P. Estevan, V. Wan, and O. Scharenborg, "Finding maximum margin segments in speech," in *Proceedings of ICASSP*, Honolulu, Hawai'i, USA, April 2007, vol. 4, pp. 937–940.