Important factors in HMM-based phonetic segmentation

D.R. van Niekerk and E. Barnard

Human Language Technologies Research Group,
Meraka Institute, Pretoria / North-West University, Potchefstroom

dvniekerk@csir.co.za, ebarnard@csir.co.za

Abstract

When doing research into or building systems involving spoken language, one invariably relies on relevantly annotated speech data for analysis and incorporation into such systems. We investigate methods and parameters for a baseline phonetic segmentation system on a few South African languages with the intention of determining how accurately we can apply basic methods and characterising typical deficiencies with the goal of defining further refinement strategies. An HMM-based system with a single mixture per triphone is found to work well, though the accurate segmentation of plosives remains a challenge. Suggestions for addressing this challenge are presented.

1. Introduction

Modern techniques for the development of spoken language technologies such as speech recognition and synthesis are reliant on large sets of speech data in the form of annotated audio recordings. For purposes such as speech synthesis or data modelling, these annotated corpora are typically described by labels that define the temporal locations of phones, which represent the acoustic realisations of the smallest meaningful units of speech, namely phonemes. Data in this form can be used to construct language based systems (including speech recognition and synthesis systems) through the training of statistical models or the definition of acoustic databases, as well as aid language research in general.

The accuracy and consistency of phonetic labels are crucial to the eventual quality of systems dependent on speech data. Labels that accurately isolate phones in training data used in statistical models such as Hidden Markov Models (HMMs) are useful as bootstrap data aiding in the initialisation of these models. This can have significant positive effects on the performance of speech recognition systems [1, 2, 3]. Other applications that use labeled corpora in a more direct form, such as concatenative speech synthesis systems, where an acoustic segment inventory is compiled, also require highly accurate and consistent boundary placement between phones in order to achieve acceptable quality output [2, 4].

Despite numerous attempts at developing accurate automatic segmentation techniques described in the literature, manual segmentation is still a popular solution when building spoken language systems of high quality. This is problematic in that it severely impedes the process of building new systems due to the time consuming and expensive nature of manual segmentation [5]. It is thus prudent to consider how existing automatic methods can be further extended or improved to enhance their performance and applicability, especially under non-ideal circumstances (which is often the case in the developing world) such as limited data, speaker and language idiosyncrasies as well as suitability for differing applications (e.g. those mentioned above).

We aim to develop an accurate segmentation procedure with application in various spoken language systems and research in the South African context, by evaluating the applicability and feasibility of current methods and investigating improvements in design that could increase performance and robustness.

A popular approach to achieving highly accurate segmentation is to imitate the expert human labeling procedure where a labeler places approximate boundary locations first and subsequently refines these boundaries by taking into account more of the available features of the waveform or applying certain conventions for labeling specific classes of boundaries [6, 7]. This has led to a two stage design whereby an algorithm/model is first used to isolate phone locations (with approximate boundaries) and a subsequent algorithm/model is employed to update these initial boundary locations.

The first stage (boundary estimation) has traditionally been attempted with ideas adopted from the speech recognition field, including the Dynamic Time Warping algorithm (DTW) and the Viterbi algorithm (using HMMs) [8, 9, 10, 11].

In this paper we investigate methods and parameters for an appropriate boundary estimation stage on a number of manually labeled corpora (designed for the building of concatenative TTS systems in local languages).

1.1. Choosing a system for boundary estimation

For the purposes of boundary estimation, the popular DTW and HMM based methods were considered, both of which are implemented in *Festvox* [12]. These methods are suitable for segmenting speech with known orthographic transcriptions. DTW based segmentation is considered more accurate [10, 11], but requires the existence of a signal with similar acoustic properties (to the signal to be segmented) of which the phone boundaries are known. This is generally achieved by synthesizing a signal using the relevant transcriptions, but can be problematic when having to segment speech in a new language which is acoustically or phonetically very dissimilar to existing languages for which TTS systems exist. Although HMM based segmentation lacks accuracy when compared to DTW under ideal conditions, it is considered to be more robust in that mostly fine errors occur during segmentation as opposed to large errors in boundary placement which occur more often with DTW alignment [10, 11].

To investigate these claims, we experimented with the segmentation of South African English speech data by a female speaker, using the existing systems in the *Festvox* package. In this experiment we counted the number of labeling errors that caused the corresponding diphone to be unusable in a concatenative speech synthesis system. We found the HMM based segmentation to make no such serious errors, while segments from the DTW process contained errors making up more than 1% of all the segments considered (20300 in total), with as much as a 3% catastrophic error on some phones. Although this is a rather crude comparison, the situation tested, that of segmenting English speech is considered the most ideal practically possible scenario for DTW alignment, because of the existence of a female English voice that was used to generate the synthetic signal. Segmentation quality of speech in different languages is expected to be even worse as a result of acoustic and phonetic dissimilarities. Informal listening tests conducted on speech synthesis systems built using the two sets of alignments also indicated that the segments (containing fine errors) identified by the HMM-based system is preferable over the DTW-based alignments. Since text for recording speech databases in developing-world languages is often carefully chosen for optimal diphone coverage (in order to minimise database size and hence the costs involved in this process), even a 1% loss in diphones is sometimes unacceptable. We thus chose to continue further research into boundary estimation employing HMM-based methods.

2. Experimental setup

In order to determine how HMM-based techniques can be successfully applied to boundary estimation, a phone recogniser was implemented using *HTK* [3] and integrated into the *Festvox* voice building framework.

2.1. System design

The HMM training procedure follows conventions for building speech recognition systems described in [3], including the possibility of context dependent phones (triphones) and splitting of Gaussian mixtures. Segmentation is achieved by firstly training models, using all the speech data and available transcriptions and subsequently applying these models in forced alignment with the transcriptions in order to determine phone boundaries. The essential components and functions of the segmentation system is depicted in Figure 1.

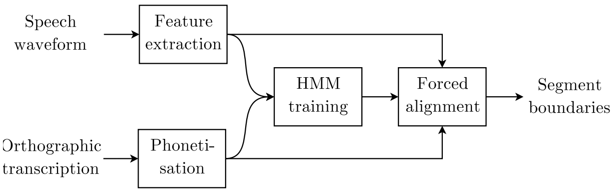

Figure 1: Segmentation system

2.2. Speech corpora

For the purpose of testing the above system, we applied it to three sets of recordings in three South African languages, namely Afrikaans, isiZulu and Setswana. These recordings were manually labeled with the aim of building concatenative TTS systems. We use these manual labels as a reference with which to compare the automatically obtained segments for various sets of parameters. The data set sizes are small, but are typical of the sizes that are currently being used to develop TTS systems for local languages [13], using careful text selection strategies to ensure diphone coverage (Table 1).

Language	Gender	Utterances.	Duration	Phones
Afrikaans	Male	134	21 mins.	12336
isiZulu	Male	150	19 mins.	8569
Setswana	Female	332	44 mins.	26009

Table 1: Reference data sets

2.3. Measures of comparison

Considering the fact that we are interested in the quality of segmentation for general purposes, we employ two measures of comparison between automatic and reference labels. Firstly, the traditional boundary accuracy (where boundaries falling within a certain threshold of the reference are counted as correct) and secondly the “Overlap Rate” (O.R.), which involves calculating how much segments overlap in time, in a duration-independent way [10]. Briefly, the O.R. is given by:

$$O.R. = \frac{C\_Dur}{M\_Dur} \quad (1)$$

$$= \frac{C\_Dur}{R\_Dur + A\_Dur - C\_Dur} \quad (2)$$

where C_Dur, M_Dur, R_Dur and A_Dur are the *common*, *maximum*, *reference* and *automatic* durations respectively (see Figure 2).

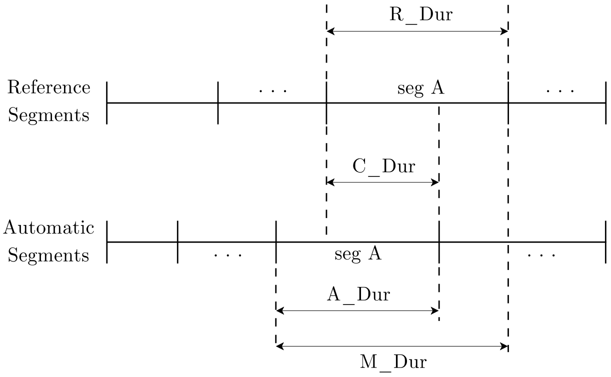

Figure 2: “Overlap Rate” definition [10]

It is important to note here that the phonetic sequence is known and as such, it is known exactly which reference segment to compare with a particular automatic segment. Thus even when no overlap occurs or when multiple segments overlap with incorrect reference or automatic segments, this merely results in $C\_Dur = 0$ and thus $O.R. = 0$.

With respect to the traditional boundary accuracy measure, we used the conventional 20ms threshold throughout.

2.4. Factors in HMM-based segmentation

Using the above system, corpora and measures, it is possible to evaluate factors impacting the accuracy and robustness of segmentation. Factors of particular interest include the following:

- Feature vectors used and parameters concerning how they are extracted,

- HMM model parameters, such as context dependent or independent models, number of Gaussian mixtures and topology,
- Speaker and gender effects on accuracy,
- Segmentation performance over different languages,
- Segmentation accuracy for particular phone and boundary categories, and
- Typical problems with HMM-based segmentation.

As a starting point we select system parameters that have been proven to work well in the domain of speaker independent speech recognition [14] and use these parameters as a baseline for comparison.

3. Results

In this section we describe some typical difficulties experienced when applying HMMs to segmentation, showing how these difficulties affect segmentation accuracy. Furthermore, we present practical advice concerning operating parameters and conditions for accurate segmentation.

All results in Section 3.1 are calculated from the most successful parameters determined in Section 3.2. Thus, three state left-to-right, single Gaussian mixture, context-dependent models (triphones) trained with MFCC (Mel Frequency Cepstral Coefficient) feature vectors incorporating delta and acceleration coefficients extracted every 5ms and using a 10ms window size.

3.1. Typical errors during segmentation

3.1.1. Phone isolation

Using the "Overlap Rate" previously described, it is possible to judge how well each phone is isolated . Figure 3 shows the accuracies for particular phonetic categories. From these results, it is evident that fricatives, nasals and vowels are isolated more accurately than plosives, silences and trills. This is possibly attributable to the nature of the feature vectors used (Section 3.2.1), causing models for spectrally more distinguishable segments to be more successful. Another possibility is that the successful training of models for these "problem phones" is less easily achieved from the "flat start" approach to HMM initialisation that is commonly used.

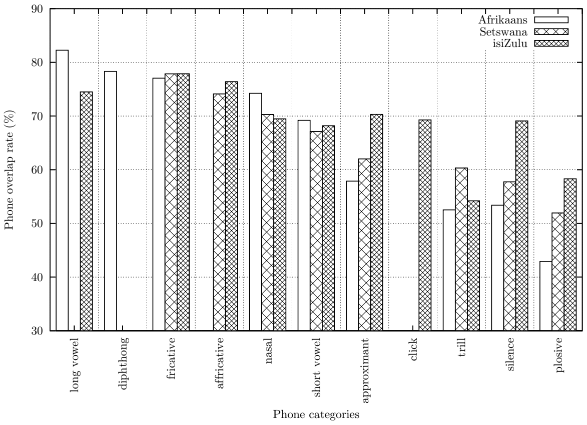

Figure 3: Overlap rates for different phonetic categories

3.1.2. Phone transitions (Boundaries)

Another important aspect of the segmentation accuracy is how close boundaries are placed to the actual transitions between phones. By looking at boundary placement accuracies, we have identified problematic cases. Some of the results are not surprising when one considers that the nature of certain transitions cause difficulties even for human labelers. These difficulties are most often overcome during manual labeling by simply defining clear conventions which determine where boundaries are placed when the transition cannot be easily perceived. Table 2 shows some examples of transitions that are of particular interest.

Transition	Afrikaans	Setswana	isiZulu
short vowel - short vowel	47,06%	47,99%	43,75%
nasal - nasal	NA	19,19%	NA
closure - plosive	59.62%	62.51%	82,79%

Table 2: Boundary accuracies for classes of phone transitions that are not segmented well

It is also interesting to note that there are some transitions which are considered easy by human labelers that are not always well placed (e.g. transition between closures and plosives). In contrast with these problematic cases, there are also some transitions which are consistently accurately identified (Table 3).

Transition	Afrikaans	Setswana	isiZulu
fricative - short vowel	90,58%	86,46%	90.07%
silence - short vowel	89,74%	73,72%	95,92%

Table 3: Boundary accuracies for classes of phone transitions that are segmented well

3.2. Practical segmentation parameters

In this section we present results to experiments that were performed with parametric variation of the segmentation process. We focused on the feature extraction and HMM training procedures (see Figure 1) as well as an analysis of the errors made by the system.

3.2.1. Feature vectors

During the feature-extraction phase, the speech signal is parametrised into a sequence of feature vectors by windowing the signal at regular intervals and calculating a relevant representation for each window. *HTK* provides a number of options for specifying how parametrisation is done, including feature vector type, window and step sizes.

Taking into account that the segmentation system trains and applies the HMM models on a single speaker only, our first concern was the applicability of the window and step sizes that are commonly used for speech recognition purposes. Window sizes of around 20 to 25ms with a step size of 10ms have been shown to work well for speaker-independent speech recognition, but are not necessarily optimal for our purposes. Figures 4 and 5 show the results obtained when stepping through various window sizes. In each case the step size is equal to half the window size so that consecutive windows overlap 50% in time.

These results seemed to indicate that all of the data sets benefited from higher resolution feature extraction. Table 4 (on the final page) shows why this could be the case. Our conventional three-state, left-to-right HMM topology imposes a min-

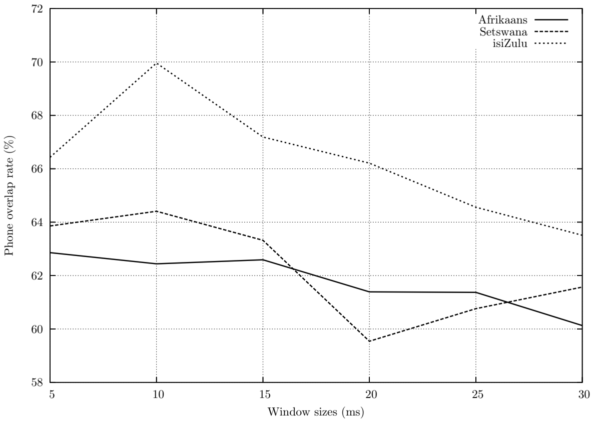

Figure 4: Phone overlap rates for the three languages using varying window/step sizes

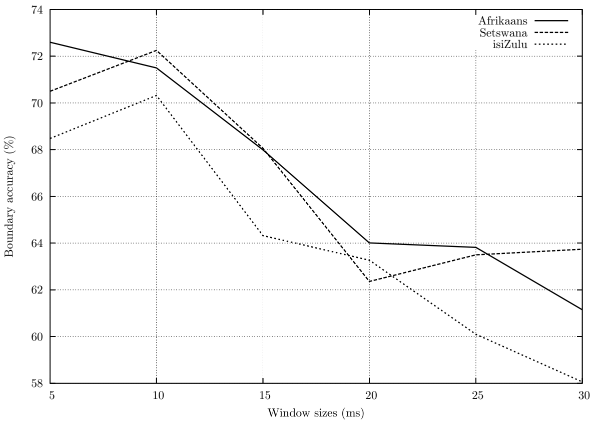

Figure 5: Boundary accuracies for the three languages using varying window/step sizes

imum phone duration constraint of 3 times the step size (that is a minimum phone duration of 30ms for commonly used step sizes). This is obviously not appropriate for all phone types.

We also experimented with different feature representations including Linear Prediction Coefficients (LPCs) and Mel Frequency Cepstral Coefficients and found MFCCs with delta and acceleration coefficients to give much better performance (this is consistent with literature on speech recognition [14]). Techniques like Cepstral Mean Normalisation (CMN) had little effect on the outcome. This is a sensible result when one considers that all models are trained and applied on a single voice and channel.

3.2.2. Models

Concerning the training of HMM models, we firstly evaluated the techniques and factors classically used in speech recognition, to ascertain the validity of these methods in the segmentation domain. These include:

- Effects of training data set size on accuracy,
- Context independent and context dependent HMMs, and
- Number of Gaussian mixtures per state.

It is interesting to note that some of the factors that have great impact on speech-recognition performance (especially for the speaker independent case), have smaller effects in the segmentation scenario (training and applying HMMs on the same set of data by only a single speaker and defining phone overlap as performance measure). The results of our investigations are briefly discussed below.

Size of the data set

An important factor when training statistical models is the size of the training set. For speaker-independent speech recognition, a significant number of training samples of each phone is necessary to obtain accurate recognition rates. In our scenario, we were able to get decent segmentation accuracy even with very little data (Table 5). Segmentation consistency did, however, gradually improve with the addition of more training samples.

Set size	Afrikaans		Setswana		isiZulu	
	O.R.	total time	O.R.	total time	O.R.	total time
large	NA	NA	64.41	2642s	NA	NA
medium	62.44	1234s	64.05	1347s	69.96	1132s
small	63.05	643s	65.09	625s	69.82	569s
very small	59.50	336s	62.46	284s	65.60	273s

Table 5: Phone overlap rates for different data set sizes

Context dependent phones

The use of context-dependent phones (i.e. triphones) is a common technique in speech-recognition systems to build more accurate models of phonetic segments. This is because the spectral features can vary greatly depending on the context in which a phone was realized. The use of triphone models for segmentation proved to be slightly more accurate and resulted in greater overlap consistency than monophones (Table 6), despite the small amount of available training data.

Models	Afrikaans		Setswana		isiZulu	
	O.R.	σ	O.R.	σ	O.R.	σ
Monophones	60.58	24.66	61.59	24.61	65.49	23.60
Triphones	61.80	23.86	62.24	23.93	66.31	22.69

Table 6: Comparative phone overlap rates for context dependent and context independent models (average O.R. over differing window/step sizes)

Gaussian mixtures per state

When one trains HMMs for speaker independent speech recognition, it is usually beneficial to perform splitting of Gaussian mixtures per state in order to better model diverse qualities in phone realisation by different speakers. Figure 6 shows that mixture splitting does not hold any benefits in our context and only leads to overly complex models which reduce the speed of the training and segmentation. This is not unexpected for a speaker-specific context-dependent model (which might show a lack of diversity that could benefit from multiple mixtures), but emphasizes the contrasting requirements of speech recognition and segmentation.

Category	Afrikaans		Setswana		isiZulu	
	< 30ms	total	< 30ms	total	< 30ms	total
plosives	51.1%	1923	42.7%	2433	48.8%	762
short silences	13.5%	2313	9.3%	3670	31.4%	1156
trills	18.6%	706	7.4%	499	4.8%	21
approximants	5.0%	480	6.1%	2208	1.4%	691
fricatives	2.5%	1634	0.1%	1778	0.0%	622
nasals	0.5%	1049	0.0%	2276	0.4%	907
vowels	0.5%	3863	0.5%	10832	0.1%	3137
silences	0.0%	368	0.1%	743	0.0%	878
affricates	NA	0	0.4%	1335	0.0%	69
aspirated stops	NA	0	3.0%	235	4.7%	254
clicks	NA	0	NA	0	2.8%	72
TOTAL	12.3%	12336	6.2%	26009	8.9%	8569

Table 4: Proportions of phones with durations of less than 30ms

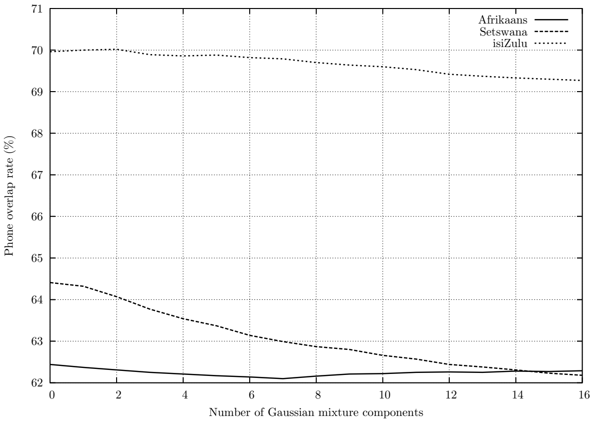

Figure 6: Phone overlap rates for the three languages for increasing number of Gaussian mixtures per state

HMM topology
In view of the phone durations summarized in Table 4, the definition of suitable HMM topologies that better suit certain phone types might present a sensible option. In a few trial runs we were able to increase the phone overlap rate by defining two-state HMMs for phones with shorter durations; however, to obtain consistent improvements it is necessary to elucidate all the important factors during training and segmentation.

4. Conclusion

With the focus on defining a system to serve as a basis for refinement and research towards robust and accurate phonetic segmentation, we investigated the feasibility of typical methods for boundary estimation. The advantage in robustness of HMM-based methods over DTW was verified and led to the further examination of important factors when segmenting speech with HMMs.

A number of interesting results are presented here with regards to segmentation accuracy under various practical conditions. Amongst others, it was shown that performance trends tended to stay consistent across languages and speakers, especially gender (refer to Figures 4, 5 and 6 as well as Table 6). This makes it possible to define practical parameters that are suitable for segmentation in the general case presented here. The increase in segmentation accuracy associated with extracting features at a higher time resolution and the identification of phone types which benefit from this is an important result that can be used to good effect during segmentation and further research in this area (e.g. by defining appropriate HMM topologies). From the results obtained for training data sets of different sizes, it is evident that robust segmentation can be achieved even on very small amounts of available data (somewhat smaller than what is generally deemed useful for building language based systems and spoken language research).

In the literature, accurate segmentation has been achieved by using elaborate statistical methods such as explicit boundary models [6, 15] and multiple segmentation machines with statistical candidate selection [4]. These techniques require large amounts of data and – more importantly – significant amounts of manually labeled data for training. We believe that a more appropriate and indeed cost-effective approach toward boundary refinement (at least in the South African context where data scarcity is a problem) involves using the known phonetic sequence and the signal properties of specific phonetic categories. This could be done by optimising the HMM-based procedure, applying boundary correction strategies (using features such as signal energy and fundamental frequency more directly) or applying "convention" rules in much the same way that human labelers do. In this paper we have shown that cases that would benefit from such strategies do indeed exist and we plan to continue research in this area.

5. References

[1] T. Laureys, K. Demuynck, J. Duchateau, and P. Wambacq, "An Improved Algorithm for the Automatic Segmentation of Speech Corpora," in *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, May 2002, vol. 5, pp. 1564–1567.

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

[3] 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.

[4] S.S. Park, J.W. Shin, and N.S. Kim, "Automatic speech segmentation with multiple statistical models," in *INTERSPEECH*, September 2006.

[5] M. Wagner, "Automatic labelling of continuous speech with a given phonetic transcription using dynamic programming algorithms," in *Proceedings of ICASSP*, 1981, vol. 1, pp. 1156–1159.

[6] A. Sethy and S. Narayanan, "Refined Speech Segmentation for Concatenative Speech Synthesis," in *Proceedings of ICSLP*, 2002.

[7] E. Barnard and M. Davel, "Automatic error detection in alignments for speech synthesis," in *Proceedings of PRASA*, 2006, pp. 53–56.

[8] H. Ney, "The use of a one-stage dynamic programming algorithm for connected word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 263–271, 1984.

[9] F. Malfrère and T. Dutoit, "High-quality Speech Synthesis for Phonetic Segmentation," in *EUROSPEECH*, 1997, pp. 2631–2634.

[10] S. Paulo and L.C. Oliveira, *Advances in Natural Language Processing*, Springer Berlin / Heidelberg, http://www.l2f.inesc-id.pt, 2004.

[11] J. Kominek, C.L. Bennett, and A.W. Black, "Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis," in *EUROSPEECH*, September 2003, pp. 313–316.

[12] A. W. Black and K. Lenzo, *Building Synthetic Voices*, http://www.festvox.org/bsv, 2007.

[13] J.A. Louw, M. Davel, and E. Barnard, "A general-purpose IsiZulu speech synthesizer," *South African journal of African languages*, vol. 2, pp. 1–9, 2006.

[14] E. Gouws, K. Wolvaardt, N. Kleynhans, and E. Barnard, "Appropriate baseline values for HMM-based speech recognition," in *Proceedings of PRASA*, November 2004, pp. 169–172.

[15] 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.