Segmentation of speech: Child's play?

Odette Scharenborg[1], Mirjam Ernestus[2,3], Vincent Wan[4]

[1]Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
[2]Department of Linguistics, Radboud University Nijmegen, The Netherlands
[3]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[4]Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, UK
O.Scharenborg@let.ru.nl, Mirjam.Ernestus@mpi.nl, V.Wan@dcs.shef.ac.uk

ABSTRACT

The difficulty of the task of segmenting a speech signal into its words is immediately clear when listening to a foreign language; it is much harder to segment the signal into its words, since the words of the language are unknown. Infants are faced with the same task when learning their first language.

This study provides a better understanding of the task that infants face while learning their native language. We employed an automatic algorithm on the task of speech segmentation without prior knowledge of the labels of the phonemes. An analysis of the boundaries erroneously placed *inside* a phoneme showed that the algorithm consistently placed additional boundaries in phonemes in which acoustic changes occur. These acoustic changes may be as great as the transition from the closure to the burst of a plosive or as subtle as the formant transitions in low or back vowels. Moreover, we found that glottal vibration may attenuate the relevance of acoustic changes within obstruents. An interesting question for further research is how infants learn to overcome the natural tendency to segment these 'dynamic' phonemes.

Index terms: unsupervised speech segmentation, speech analysis, articulatory features, infant language acquisition.

1. INTRODUCTION

A speech signal does not contain (many) obvious markers – like the space between words in a written text – to indicate word boundaries. A person listening to the speech signal is thus faced with the task of segmenting the speech signal into words in order to obtain the message in the speech signal. The difficulty of the word segmentation task is immediately clear when listening to a foreign language. Whereas the speech signal in one's own mother tongue is easily segmented into words, the segmentation of the speech signal of a foreign language is much harder – if not impossible – since the words that constitute the language are unknown.

The latter situation is exactly the circumstances under which infants have to learn to speak and understand their native language. Psycholinguists have found that young infants can discriminate among virtually all sounds used in all languages, whereas adults cannot [1]. This capability is however lost very soon. In language acquisition, infants first learn which phonetic contrasts are important in the language they are learning [1]. In a subsequent step, infants learn to group together sounds that may sound distinct, for instance, due to coarticulation or to speaker differences in gender, age, speaking style or rate, but nevertheless belong to the same 'phonetic unit' (this is called 'categorisation') [1].

Our study aims at getting a better understanding of human speech processing, by building a computational model of human speech recognition (on the basis of SpeM [2], using techniques from the field of automatic speech recognition), that is able to model all parts of the human speech recognition process, including the acquisition of new phonemes and, subsequently, words. In this context we are interested in getting a better understanding of the task infants face while learning their native language. We employed an automatic algorithm (Section 2.2, [3]) on the task of unsupervised *speech* segmentation.

The algorithm was tuned such that the number of hypothesised boundaries was equal to the number of boundaries in our reference transcription. We then assessed the performance of our speech segmentation algorithm by comparing the boundaries hypothesised by the algorithm to the reference phoneme boundaries. In addition to the correctly hypothesised boundaries (and the boundaries that were erroneously missed, see also Section 2.3), the algorithm also hypothesised boundaries that are *not* in between two phonemes (i.e., *not* located on phoneme boundaries). Infants eventually learn which acoustic events belong together in a phonetic unit. At first sight, automatic algorithms seem to have difficulty with this task since they hypothesise boundaries *inside* phonemes. In this paper, we analyse these boundaries that are hypothesised inside a phoneme, and try to predict where to expect these additional boundaries. This enterprise provides us on the one hand with more insights into the difficulty of the task infants face while learning their native language, and on the other hand gives us indications how to improve our speech segmentation algorithm.

2. EXPERIMENTAL SET-UP

2.1. Material

In this study, the TIMIT [4] speech corpus was used. It consists of reliably hand labelled and segmented data of quasi-phonetically balanced sentences read by 630 native speakers of eight major dialect regions of American English. Of the 630 speakers in the corpus, 438 (70%) were male. For the analyses, TIMIT's standard test set (excluding the `sa` sentences) was used, consisting of 1,344 utterances.

The speech was parameterised with 12 Mel Frequency Cepstral Coefficients (MFCCs) and log energy, augmented with their first and second derivatives resulting in 39-dimensional MFCC vectors. The MFCCs were computed on windows of 15 ms, with a 5 ms frame shift (the window size and frame shift were determined in a separate tuning experiment), and cepstral mean and variance normalisation was applied.

2.2. The speech segmentation algorithm

The algorithm used to segment the speech is described in [3] and relies on a method called *maximum margin clustering* (MMC) [5]. Without access to the phone labels, the algorithm segments the speech into clusters of input frames that 'belong together'. The speech segmentation algorithm is thus *un*supervised.

The objective of MMC is to find a dichotomy of a given set of unlabelled MFCC vectors (see Section 2.1) such that the margin separation between the two resultant groups is maximal (see Figure 1). Figures 1a and 1b are examples of a non-optimal decision boundary. The empty region bounded by the two lines is called the margin and should have maximal width, that is, it should be as wide as possible while remaining empty. The MMC extends this principle to the non-separable case by penalising incursions into a so-called *soft-margin* and the goal then is to maximise the soft-margin while minimising the penalties. Figure 1c is an example of an optimal decision boundary.

Using a sliding window 18 MFCC vectors wide (determined in a separate tuning experiment), a set dichotomy is obtained for the frames inside the window. In a subsequent step, the dichotomy assignments between adjacent sliding windows are compared and where a maximum margin dichotomy is consistently detected a boundary is hypothesised (for more information, see [3]).

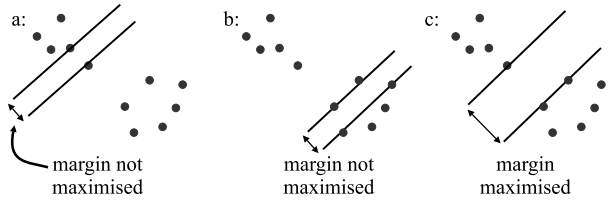

Figure 1. The maximum margin criterion.

2.3. The boundaries

The speech segmentation algorithm hypothesised 44,885 boundaries. Following the method described in [3], we defined a boundary as correctly hypothesised if it fell within a distance of 20ms from the phoneme boundary in TIMIT. This resulted in 67.9% correctly hypothesised boundaries. This result is comparable with existing unsupervised methods for automatic phoneme boundary detection (e.g., [6]; see [3] for a discussion of the results).The algorithm also hypothesised boundaries that do not coincide with the phoneme boundaries in TIMIT. Of these 13,959 additional boundaries, 3,075 boundaries were hypothesised in the silence part at the start or end of a file, 10,884 are *inside* a phoneme. In these cases, there is apparently a difference between clusters of frames inside the sliding window that is big enough to warrant hypothesising a boundary, even while there is no phonetic boundary.

3. ANALYSING THE BOUNDARIES

Since we are interested in identifying the segments that are liable to get spurious boundaries, we analysed the speech signal in terms of phonemes. In order to be able to generalise over different phonemes, we characterised phonemes by 'articulatory features' (AFs). AFs describe properties of speech production and are physiologically motivated classes which characterise the essential aspects of articulatory properties of speech sounds for speech perception (e.g., voice, nasality) [7]. In our analysis, we focused on the 10,884 boundaries that were hypothesised inside a phoneme.

3.1. The articulatory features

For the analysis, we used the set of seven articulatory features shown in Table 1. The names of the AFs are self-explanatory, except maybe for *staticity*, which states whether an acoustic change occurs (as, e.g., is the case for diphthongs; 'dynamic'), or not ('static'). The set is based on the six AFs proposed in [8], the difference being that in the present study the place of articulation for vowels and consonants have been separated into two different AFs (i.e., *place* for consonants and *height* for vowels). In TIMIT, the silence (i.e., the closure) and release (i.e., the burst) part of plosives have been annotated separately, but in our study the silence part is merged with the release part to form a single segment.

Table 1. Specification of the AFs and their respective values.

AF	**Values**
manner	approximant, retroflex, fricative, nasal, stop, vowel, silence
place	bilabial, labiodental, dental, alveolar, velar, nil, silence
voice	voiced, unvoiced
height	high, mid, low, nil, silence
backness	front, central, back, nil
roundness	rounded, unrounded, nil
staticity	static, dynamic

3.2. The analysis

For each boundary hypothesised inside a phoneme (thus for each boundary in the 10,884 set), we determined the label of that phoneme in the TIMIT transcription and the labels of the preceding and following neighbouring phoneme. Each phoneme label was then rewritten in terms of its AF values.

We analysed the likelihood of a boundary to be correct or to be hypothesised inside a phoneme by means of generalised linear mixed-effect models using the binomial link function. We used contrast coding[1], and entered the phoneme itself and the preceding and following phonemes as crossed random factors [10,11], as the AFs together form the phoneme labels. The AFs of the phoneme itself and of the preceding and following phonemes were entered as predictors.

A generalised model, with the binomial link function, has the form

$$\text{logit}\, p = c + \beta_1 + \beta_2 + \beta_3 + \ldots,$$

where logit p represents log $[p(1\text{-}p)]$, and p is in our case the probability of a boundary to be hypothesised inside a phoneme. The constant c is the intercept, and represents the default phoneme. The different βs [11] represent the relevance of the different AFs for the estimation of the logit p, and were estimated with maximum likelihood.

In the following analyses, only those effects are reported that are statistically significant (calculated using F-tests). In addition, we report the absolute estimated values of the different βs, with an explanation of whether the likelihood of additional boundaries increases or decreases for each effect.

[1] One phoneme or combination of AF values is used as the 'Intercept', i.e., the default, with which all other phonemes or combinations of AF values are compared.

Table 2. Percentage correctly hypothesised boundaries at the end of a phoneme of 'Total' number of phonemes with the specified manner AF value.

AF value	%	Total
'vowel'	72.3	14,888
'stop'	69.5	6,341
'fricative'	68.9	7,832
'nasal'	66.4	4,434
'retroflex'	59.2	3,541
'approximant'	56.0	3,480
'silence'	27.3	6,353

4. RESULTS AND DISCUSSION

Since only *manner* can be meaningfully specified for all phonemes, we first analysed all data with only the manners of the phoneme itself and of the preceding and following phonemes as predictors. We observed robust effects of the *manner* of the phoneme itself ($F(6,40159)=2.9212$, $p < 0.01$) and of the preceding phoneme ($F(6,40159)=18.3521$, $p < 0.001$). The *manner* of articulation of the following segment appeared not to have an effect.

Additional boundaries were more likely in vowels than in nasals ($\beta=1.279202$, $p < 0.001$). This difference can most likely be attributed to the *staticity* of the phoneme (see also below). During the realisation of a diphthong vowel the articulators move from one position to the next, resulting in acoustic change. Since the algorithm is designed to group together frames that are similar, the acoustic change results in the hypothesis of (additional) boundaries, and the algorithm divides the diphthong into two separate segments. Nasals, on the other hand, are more or less 'static' sounds, and this staticity results in less additionally hypothesised boundaries.

Additional boundaries were also more likely after vowels than after fricatives ($\beta=0.264826$, $p < 0.001$). Moreover, additional boundaries were more likely after any segment than after a silence (βs range from 0.547408, for fricatives, to 0.813009, for vowels, all ps < 0.001). Interestingly, these results show exactly the opposite pattern as the percentages of correctly hypothesised boundaries at the end of phonemes and silences, listed in Table 2. Boundaries indicating the end of 'vowel' segments are typically hypothesised fairly well, whereas the end of a 'silence' segment tends to be hypothesised poorly. The latter effect is most likely due to the endpointing algorithm which is used to remove the silence at the beginning and end of each utterance (note that the 'Total' number for 'silence' in Table 2 includes the silences at the beginning and end of each utterance), which does not only lead to missing boundaries at the end of the silences, but also to fewer additional boundaries within the directly following segments.

In order to test the role of the other AFs in the hypothesising of additional boundaries, we analysed obstruents, nasal consonants, and vowels separately, and investigated which of their characteristics predict the presence of an additional boundary. Since *voice* is a meaningful specification only for plosives and fricatives (vowels and nasals are always 'voiced'), we grouped the plosives and fricatives together and analysed this category of obstruents. For the obstruents, *voice*, *manner* (thus either 'stop' or 'fricative'), *staticity*, and *place* of articulation are meaningful AFs. Only *voice* showed a main effect ($F(1,14076)= 7.3711$, $p < 0.01$, $\beta=0.9136$): additional boundaries were more likely in 'unvoiced' (3,269 additional boundaries versus 5,978 expected boundaries) than in 'voiced' obstruents (1,010 versus 3,821).

This result may be somewhat surprising since *staticity* may be expected to be a better predictor than any other AF, including *voice*. In our description of the phonemes, most fricatives are described as 'static' and all stop consonants as 'dynamic' (remember that the closure and release part of plosives are labelled as one segment – contrary to the standard TIMIT labelling), and distinction is made between 'voiced' and 'unvoiced' obstruents, even though 'unvoiced' obstruents are obviously more dynamic than 'voiced' ones. Our results show that the distinction between 'voiced' and 'unvoiced' obstruents is highly important and suggest that the AF *staticity* should be made sensitive to *voice*.

For nasal consonants, we studied the role of *place* of articulation and *staticity*. Both predictors emerged as significant (*place*: $F(2,3314)=3.2801$, $p < 0.05$; *staticity*: $F(1,3314)=28.6835$, $p < 0.001$). Additional boundaries were more likely in the 'bilabial' ([m] and syllabic [m], $\beta=0.3615$) and 'velar' ([ŋ] and syllabic [ŋ], $\beta=0.5981$) nasals than in the 'alveolar' ones ([n] and syllabic [n], all ps < 0.001). Phonetic research [12] has shown that 'alveolar' nasals (like /n/) are often partially assimilated to the following phoneme. For instance, an /n/ followed by a bilabial stop (e.g., [b]) is often realised as an [n] that gradually becomes more bilabial ([m]) like. As a consequence, the formant transitions at the end of 'alveolar' nasals are not as great as those at the end of non-alveolar nasals. As explained before, dynamic change is the basis for hypothesising boundaries.

Additional boundaries were also more likely in 'dynamic' than in 'static' nasals ($\beta=0.9909$). The 'dynamic' nasals are the syllabic nasals. We expect formant changes to be greater in syllabic than in non-syllabic nasals, as they form the transitions from consonants to consonants, instead of from vowels to other vowels. This explains the increased number of additional boundaries.

Vowels differ in their specification for *height*, *backness*, *roundness*, and *staticity*. Three predictors appeared significant: *height* ($F(2,14623)=26.946$, $p < 0.001$), *backness* ($F(2,14623)= 22.001$, $p < 0.001$), and *staticity* ($F(1,14623)=23.328$, $p < 0.001$). Additional boundaries were more often positioned in 'low' than in 'high' ($\beta=1.0332$) vowels, more often in 'back' than in 'central' ($\beta=1.3971$) and 'front' ($\beta=0.3425$) vowels, and more often in 'dynamic' than in 'static' ($\beta=0.6544$) vowels (all ps < 0.01). The higher number of additional boundaries for 'dynamic' is again according to expectation.

The higher number of hypothesised boundaries for 'back' compared to 'central' can also easily be explained. The formant transitions in a 'central' vowel are much smaller (thus less acoustic change) than the formant transitions in a 'back' or 'front' vowel. It is thus to be expected that more additional boundaries are hypothesised in 'front' and 'back' vowels than in 'central' vowels, and that the difference between 'back' and 'front' vowels is smaller than the difference between 'back' and 'central' vowels, which is exactly what we observed.

Additional boundaries were more often hypothesised in 'low' vowels than in 'high' vowels. During the production of a 'low' vowel, the mouth is much more open than during the production of a 'high' vowel. During the production of the constriction of the preceding and following consonants, the mouth needs also to be fairly closed. As a consequence, the formant transitions are comparatively greater in 'low' vowels, which in turn implies more acoustic change and thus more additionally hypothesised boundaries than in 'high' vowels.

We then investigated the role of the characteristics of the preceding obstruents, nasal consonants, and vowels. If the preceding

segment was an obstruent, we found no effects for its *place*, *manner*, or *voice* (all ps > 0.05). Preceding nasals showed an effect of *place* of articulation ($F(2,4110)=4.3651$, $p < 0.05$). The bilabial [m] and syllabic [m] were less likely to be followed by a segment with an additional boundary than the 'alveolar' ($\beta=0.18930$) and the 'velar' ($\beta=0.39331$, both ps < 0.05) nasals. This pattern of results is exactly the opposite as the one we found for additional boundaries *within* nasals. We do not have an explanation for this at this moment. For preceding vowels, we found an effect of their *height* ($F(2,13296)=4.4611$, $p < 0.05$) and *staticity* ($F(1,13296)=11.5797$, $p < 0.001$). Additional boundaries were more likely after 'high' than after 'low' vowels ($\beta=0.18644$), and, again according to expectation, more likely after 'dynamic' than after 'static' vowels ($\beta=0.16017$). Interestingly, 74.5% of the boundaries at the end of a 'high' vowel were correctly hypothesised, while only 66.8% of the boundaries at the end of a 'low' vowel were correctly hypothesised. Apparently, 'high' vowels more often lead to additional boundaries in the following segments, but less often have additional boundaries in the vowels themselves (see above).

Finally, we tested the predictive power of the characteristics of following obstruents, nasal consonants, and vowels. We did not find any effects for following obstruents and vowels. With respect to nasal consonants, we only found a significant difference between 'alveolar' and 'bilabial' nasal consonants ($F(2,4167)=3.1493$, $p < 0.05$). Additional boundaries were more likely in segments preceding [n] and syllabic [n] ($\beta=0.53086$). It is well-known that vowels may become nasalised before nasals. We therefore expected an effect of 'nasal' on the number of additionally hypothesised boundaries. Indeed, when it is known that the following phoneme is a 'nasal', more boundaries are being hypothesised in the current phoneme.

5. CONCLUDING REMARKS AND FUTURE WORK

This paper tries to get more insights into the difficulty of the task infants face when learning their native language. We employed an automatic unsupervised algorithm on the task of segmenting the speech into parts, and subsequently analysed the boundaries that were erroneously hypothesised *inside* a phoneme in order to ascertain which characteristics of the speech signal are responsible for these erroneous boundaries. In short, the algorithm consistently placed additional boundaries in phonemes in which acoustic changes occur. These acoustic changes may be as great as the transition from the closure to the burst of a plosive, but also as subtle as the formant transitions in low or back vowels. Moreover, we found that glottal vibration may attenuate the relevance of acoustic changes within obstruents. In subsequent research, we also plan to analyse the boundaries that were *missed* by the automatic algorithm to get more insights into the difficulty of the task infants face when learning their native language.

An interesting question for further research is how infants learn to overcome the natural tendency to segment dynamic phonemes. Possibly, this is based on statistical learning: Infants discover at a certain moment that bursts are always preceded by silence and as a consequence group the silences and bursts together. We plan to test this hypothesis using our automatic speech segmentation algorithm in the near future. Another possibility is that infants take advantage of more fine grained information present in the acoustic signal.

In this study, we used MFCCs as input to the segmentation algorithm. It is however well-known that MFCCs do not capture all relevant information in the speech signal. In future research, we plan to investigate whether acoustic features that are based on rate maps [13] (which are based on knowledge of the auditory system), will improve performance.

Finally, the analyses showed that the end of a 'silence' segment tends to be hypothesised poorly. This effect is most likely due to the endpointing algorithm which is used to remove the silence at the beginning and end of each utterance. Obviously, future research is necessary to test these hypotheses.

6. ACKNOWLEDGEMENTS

This research was supported by a Veni-grant from the Netherlands Organization for Scientific Research (NWO) to the first author and by a EURYI-award from the European Science Foundation to the second author. Vincent Wan was partly supported by the EU 6th FWP IST Integrated Project AMIDA. The authors would like to thank Lou Boves for useful comments on an earlier version of this paper.

7. REFERENCES

[1] P.K. Kuhl (2004). Early language acquisition: Cracking the speech code. *Nature Reviews – Neuroscience*, *5*, 831-843.

[2] O. Scharenborg, D. Norris, L. ten Bosch, J.M. McQueen (2005). How should a speech recognizer work? *Cognitive Science*, *29:6*, 867-918.

[3] Y. Pereiro Estevan, V. Wan, O. Scharenborg (2007). "Finding maximum margin segments in speech," *Proc. ICASSP*, Honolulu, Hawaii.

[4] J.S. Garofolo (1988). "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database", *National Institute of Standards and Technology (NIS)*, Gaithersburgh, MD.

[5] L. Xu, J. Neufeld, B. Larson, D. Schuurmans (2004). "Maximum margin clustering," *Proc NIPS*.

[6] G. Aversano, A. Esposito, A. Esposito, and M. Marinaro, (2001). "A new text-independent method for phoneme segmentation," *Proc the 44th IEEE Midwest Symposium on Circuits and Systems*, vol. 2, pp. 516–519.

[7] K. Kirchhoff (1999). *Robust speech recognition using articulatory information*, Ph.D. thesis, University of Bielefeld.

[8] M. Wester (2003). "Syllable classification using articulatory-acoustic features," *Proc. Eurospeech*, Geneva, Switzerland, pp. 233-236.

[9] J.C. Pinheiro, D.M. Bates (2000). *Mixed-effects models in S and S-PLUS.* In series: Statistics and Computing. New York: Springer.

[10] D.M. Bates (2005). "Fitting linear mixed models R", *R News* 5, 27-30.

[11] S. Chatterjee, A.S. Hadi, B. Price (2000). *Regression analysis by example*. New York: John Wiley & Sons.

[12] D.W. Gow, Jr. (2001). "Assimilation and anticipation in continuous spoken word recognition," *Journal of Memory and Language*, *45*, 133-159.

[13] M.P. Cooke (1993). *Modelling auditory processing and organization.* Cambridge, UK: Cambridge University Press.