Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries

Odette Scharenborg[a)]
Centre for Language and Speech Technology, Radboud University Nijmegen, Erasmusplein 1, 6525 HT Nijmegen, The Netherlands

Vincent Wan
Department of Computer Science, Speech and Hearing Research Group, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, United Kingdom

Mirjam Ernestus
Center for Language Studies, Radboud University Nijmegen, and Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands

(Received 27 May 2009; revised 19 October 2009; accepted 26 November 2009)

Despite using different algorithms, most unsupervised automatic phone segmentation methods achieve similar performance in terms of percentage correct boundary detection. Nevertheless, unsupervised segmentation algorithms are not able to perfectly reproduce manually obtained reference transcriptions. This paper investigates fundamental problems for unsupervised segmentation algorithms by comparing a phone segmentation obtained using only the acoustic information present in the signal with a reference segmentation created by human transcribers. The analyses of the output of an unsupervised speech segmentation method that uses acoustic change to hypothesize boundaries showed that acoustic change is a fairly good indicator of segment boundaries: over two-thirds of the hypothesized boundaries coincide with segment boundaries. Statistical analyses showed that the errors are related to segment duration, sequences of similar segments, and inherently dynamic phones. In order to improve unsupervised automatic speech segmentation, current one-stage bottom-up segmentation methods should be expanded into two-stage segmentation methods that are able to use a mix of bottom-up information extracted from the speech signal and automatically derived top-down information. In this way, unsupervised methods can be improved while remaining flexible and language-independent.
 [DOI: 10.1121/1.3277194]

PACS number(s): 43.72.Ar, 43.72.Ne [SSN] Pages: 1084–1095

I. INTRODUCTION

Over the past few years, interest in the automatic segmentation of speech has increased. In the fields of automatic speech recognition and text-to-speech, there is a need for large amounts of reliably segmented speech data, for instance, for improving recognition and synthesis performance. Furthermore, automatic speech segmentation methods are used for the automatic phonetic analysis of large amounts of speech data (e.g., Kuperman *et al.*, 2007). In the past, speech data were segmented by hand, but with the need for and availability of ever increasing amounts of speech data the task of manual speech segmentation becomes too time-consuming and expensive. Furthermore, manual labeling and segmentation are subjective, resulting in significant differences in the transcriptions created by different expert listeners (Cucchiarini, 1993). Automatic systems, on the other hand, are consistent.

Automatic speech segmentation is the partitioning of a continuous speech signal into discrete, non-overlapping units. Generally, automatic speech segmentation methods are divided into two types. Supervised methods require *a priori* knowledge (e.g., Brugnara *et al.*, 1993; Kim and Conkie, 2002; Pellom and Hanson, 1998). Most of the supervised methods are based on forced alignment techniques starting from an orthographic transcription of the speech material. This means that the representation of the word or utterance in terms of discrete units is known (from a lexicon which includes the words' pronunciations) and pre-trained acoustic models of these units are needed for the forced alignment. The task of the segmentation algorithm is then to optimally locate the unit boundaries (Sharma and Mammone, 1996). Unsupervised methods, on the other hand, require no training data for segmenting the speech signal. Instead, they use sets of rules derived from or encoding human knowledge to segment speech. Acoustic (rate of) change (e.g., Sharma and Mammone, 1996; see for early work on unsupervised automatic speech segmentation, Bridle and Sedgwick, 1977; for more recent work, see below) is an example of prior human knowledge that is used to solve the speech segmentation task. The task for an unsupervised segmentation algorithm then is two-fold; the number of segments in the speech signal needs to be determined (this is usually determined by a parameter such as the parameter δ described in Sec. III) and the position of the boundaries on the basis of the acoustic signal needs to be determined (Sharma and Mammone, 1996).

[a)]Author to whom correspondence should be addressed. Electronic mail: o.scharenborg@let.ru.nl

There are some good reasons for using unsupervised methods. First of all, supervised methods require (extensive) training on (carefully prepared) speech material. The training material needs to be transcribed in terms of the units the algorithm is supposed to segment the speech signal into, usually phones. Furthermore, usually large amounts of training data are needed to train the supervised algorithms; however, large amounts of training data are not always easily obtained and neither are transcriptions. Unsupervised methods, on the other hand, do not require training; so, obviously no training material is needed. For each new language, speech style, dialect or accent, supervised algorithms may need to be retrained, whereas unsupervised methods are based on human knowledge and understanding of the nature of speech and are therefore language and speech style independent. Furthermore, supervised methods require the units to be defined beforehand, e.g., phones, diphones, syllables, and words, in order to be able to train models for them, whereas unsupervised methods, in principle, do not. Thus unsupervised methods yield a desirable and more flexible framework for the automatic segmentation of speech. Finally, unsupervised segmentation methods are generally simpler algorithms than supervised methods (Dusan and Rabiner, 2006).

This paper focuses on unsupervised speech segmentation. A review of the current approaches for unsupervised speech segmentation shows that although very different approaches are used, the results obtained are remarkably similar (see Sec. II). Nevertheless, unsupervised speech segmentation algorithms are not yet able to perfectly reproduce manually obtained reference transcriptions (see also Sec. II). This paper compares a phone segmentation obtained using only the acoustic information present in the signal with a reference segmentation created by human transcribers. We present an in-depth analysis (Sec. V) of the output of our unsupervised speech segmentation algorithm (see Sec. III) on the task of segmenting speech from the TIMIT database (Garofolo, 1988; Sec. IV). This unsupervised speech segmentation algorithm uses acoustic change as the criterion to segment the speech signal into phones.

Naturally, the choice of the automatic segmentation method, and the assumptions underlying the method, will have an impact on the segmentation results. We chose to analyze the results of an automatic segmentation algorithm that only uses acoustic change as the criterion for hypothesizing a segment boundary, as we believe that other criteria, such as heuristics or sophisticated signal processing, will add additional complexity to the underlying system, which will have an impact on the results. If we then perform analysis on such a highly complex automatic segmentation algorithm, the results will be highly specific to that segmentation algorithm. Additionally, more convoluted results will mean that the effects of different parts of the segmentation algorithm may become difficult to tease apart. Hence we use an automatic segmentation method that only uses acoustic change as a criterion in order to ensure that the analysis is clean. Furthermore, we believe that the assumptions underlying our method have implications for other unsupervised automatic speech segmentation methods. We will address this issue in Sec. VI.

This enterprise will indicate where acoustic change is indeed a correct indicator of a segment boundary, and where it is not, thus revealing weaknesses in the criterion of acoustic change for unsupervised automatic speech segmentation. Furthermore, since most unsupervised speech segmentation algorithms use acoustic change as the means to decide when to hypothesize a boundary we believe that the analysis presented here is of interest for unsupervised speech segmentation, in general, and will reveal weaknesses in automatic speech segmentation technologies. This paper ends with suggestions on how to develop unsupervised speech segmentation algorithms that are able to create segmentations that are closer to those created by human transcribers (Sec. VI).

II. PERFORMANCES OF UNSUPERVISED SPEECH SEGMENTATION APPROACHES: A BRIEF OVERVIEW

It is not straightforward to compare the performances of different unsupervised speech segmentation algorithms described in the literature as algorithms are often tested on different sets of speech material. To make the comparison here as fair as possible, we will only discuss those methods that have been tested on TIMIT as the majority of the reported algorithms have been tested on this speech corpus.

The performance of speech segmentation algorithms is usually assessed by comparing their segmentation to a "ground truth," which usually consists of manually placed boundaries, such as those provided with the TIMIT database. A hypothesized boundary is judged to be correctly placed if it falls within a "tolerance window" from the segment boundary in the ground truth segmentation. Generally, a tolerance window of 20 ms is used (although some researchers report performances for a range of distances). This distance of 20 ms is somewhat arbitrarily chosen; however, it is backed-up by evidence from manual segmentation by Wesenick and Kipp (1996). They found that on a set of 64 read German sentences hand segmented by three humans, the mean deviation in placement of the segment boundaries could be as large as 16 ms, while 93% of the manual segmentations were inside a 15 ms time interval and 96% within 20 ms. So, 20 ms also seems to be a window within which human segmentations are in agreement with one another. The performance of a segmentation algorithm is generally presented in terms of the correct detection rate (CDR), which is expressed as

$$\mathrm{CDR} = \frac{\#\mathrm{boundaries\_correct}}{\#\mathrm{boundaries\_truth}} \times 100, \quad (1)$$

where boundaries_correct are the hypothesized boundaries that fell within the tolerance window distance from the ground truth boundaries, and boundaries_truth are the boundaries in the ground truth segmentation.

A second important issue when comparing the performances of different segmentation methods is the number of boundaries hypothesized: with an equal number of correctly placed boundaries, the method that has a number of hypoth-

esized boundaries closest to the number of actual boundaries in the ground truth is better. This is expressed as a percentage over- (or under-)segmentation (OS):

$$\mathrm{OS} = \left(\frac{\#\text{boundaries\_found}}{\#\text{boundaries\_truth}} - 1 \right) \times 100, \quad (2)$$

where boundaries_found are the boundaries hypothesized by the segmentation algorithm, and boundaries_truth are the boundaries in the ground truth segmentation.

Pereiro Estevan *et al.* (2007) presented an unsupervised speech segmentation method based on maximum margin clustering (see Sec. III for more details). At a tolerance window of 20 ms, and an over-segmentation of −1.4% (i.e., an under-segmentation), the method obtained a CDR of 67.9% on the TIMIT test data. Aversano *et al.* (2001) obtained a CDR of 73.6% correct on 480 utterances produced by 48 speakers from the TIMIT database, at an over-segmentation of 0% and a 20 ms tolerance window, using a method that captures the changes in speech signals defined as a "jump function" and subsequently hypothesizes boundaries at the peaks of the jump function. Qiao *et al.* (2008) tried to solve the segmentation problem by searching for the "optimal segmentation" using a probabilistic framework. Their "rate distortion estimated by a full covariance matrix" method obtained a CDR of 76.7% on the training set of TIMIT, using a 20 ms tolerance window; they, however, did not report the over-segmentation rate. Dusan and Rabiner (2006) presented a method that detects boundaries by searching for the peaks in a spectral transition measure. They obtained a performance of 84.6% of correctly detected boundaries at a 20 ms tolerance window on the training set of TIMIT; they did not report over-segmentation rates.

III. THE SPEECH SEGMENTATION ALGORITHM

As explained above, we investigated the fundamental problems for automatic speech segmentation algorithms. To that end, we used a segmentation algorithm based on acoustic change as a criterion in order to ensure that the analysis is clean. We opted for the unsupervised speech segmentation algorithm presented in Pereiro Estevan *et al.*, 2007. The reasons for this choice are two-fold. First of all, in order to investigate the fundamental problems for automatic speech segmentation algorithms, patterns in the errors are more easily detected when there are more errors. Second, the algorithm by Pereiro Estevan *et al.* (2007) was easily available.

The speech segmentation algorithm relies on a method called maximum margin clustering (MMC) (Xu *et al.*, 2004). MMC is a promising kernel method. It is an unsupervised form of support vector machine (SVM) (Burges, 1998): the two are related by the maximum margin criterion for finding the optimum solution. The objective of MMC is to split a set of unlabeled feature vectors (represented by the dots in Fig. 1) such that the margin separation between the two resulting sets or clusters is maximal. Figures 1(a) and 1(b) are examples of a *non-optimal* dichotomy. 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. Figure 1(c) is an example of an optimal dichotomy.

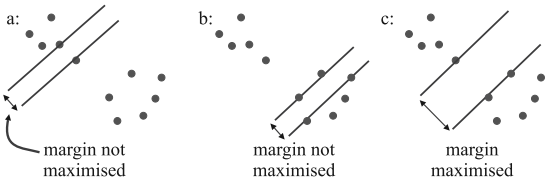

FIG. 1. The maximum margin criterion: (a) and (b) are examples of a non-optimal dichotomy, and (c) is an example of an optimal dichotomy.

The speech was parameterised by 12 mel frequency cepstral coefficients (MFCCs) and log energy and 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, and cepstral mean and variance normalization was applied.

A sliding window 18 MFCC vectors wide (90 ms) is used to isolate a small section of the signal for analysis. The MMC algorithm is applied to the MFCC vectors inside the window. Each MFCC vector is assigned to either the left or right of the margin based on the maximum margin criterion, resulting in two clusters such that the MFCC vectors on one side of the margin are more similar to one another than the MFCC vectors on the other side of the margin, and ensuring that the margin between the two clusters is maximized. Note that, in the present study, clustering ignores the time ordering of the MFCCs. The cluster assignment is plotted in the first column of Fig. 2(a) with a different shading in each element to indicate the frames' assignments, thus each element in the first column of Fig. 2(a) indicates the cluster assignment of one of the 18 MFCC vectors. The sliding window is shifted by one frame (5 ms) and the process is repeated producing the subsequent columns of Fig. 2(a). Thus, each MFCC vector is part of the analysis window multiple times (i.e., 18 times, as the window is 18 frames wide). When a boundary (a change in the shading) is hypothesized in one column, then in a subsequent column the same boundary should occur one frame earlier (i.e., lower) in the window, as the MFCC

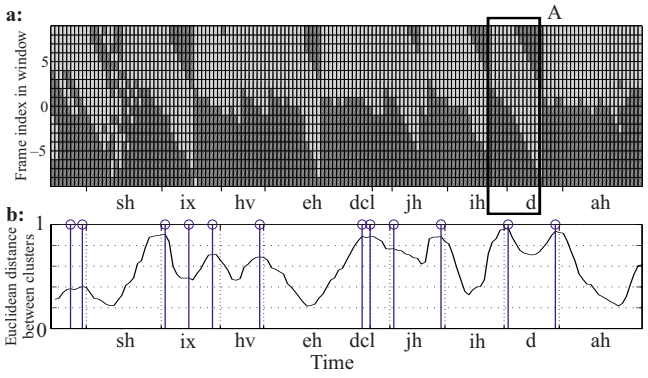

FIG. 2. (Color online) (a) Sliding window frame assignment representation; each column shows the cluster label assignments (light/dark gray) at a given time. (b) Euclidean distance between the means of the clusters; the detected boundaries are indicated by the solid vertical lines. Horizontally, the TIMIT segments and boundaries are indicated with the dashed vertical lines of the phrase "She had your dar(k)" (the [k] is not present in the figure; in IPA: /ʃɪ hɛd jɔr dɑ/).

vector's position has shifted one frame compared to its position in the previous analysis window. Hypothesized boundaries thus manifest themselves as diagonal structures in Fig. 2(a) such as shown in rectangle *A* of Fig. 2(a).

The method used to find potential boundaries consists of a combination of two approaches. The first method detects the diagonal structures in Fig. 2(a) using a mask, which is divided along its diagonal into two: each element in the upper right triangle must match the lighter shaded elements of the graph while the elements in the lower left triangle must match the darker shade. The total number of matching elements in the mask is counted each time. When all of the mask's elements are matched, then a segment boundary is marked at the middle vector. Tuning experiments (on the TIMIT test set) described in Pereiro Estevan *et al.* (2007) showed that the optimal size of the mask is a 4×3 matrix. The second approach plots the Euclidean distance between the centers (means) of the two clusters in the sliding analysis window in MFCC space. When the Euclidean distance between the cluster means shows a peak [vertical lines in Fig. 2(b)] then a boundary is hypothesized. A parameter δ controls the sensitivity of the peak detector. It specifies the minimum amount that the curve must decrease and then increase before another peak is detected. A smaller δ results in more hypothesized boundaries, whereas a greater δ results in fewer hypothesized boundaries. Tuning experiments (on the TIMIT test set) described in Pereiro Estevan *et al.* (2007) showed that $\delta=0.001$ yielded the best results.

The detected boundaries from both approaches are combined (in a left-to-right fashion) such that the resulting set of detected boundaries consists of all boundaries hypothesized by either approach. However, if both methods hypothesize boundaries within two vectors (10 ms) of each other, then they are replaced by a single boundary located at the frame halfway between the two. We refer to this as "smoothing."

IV. THE SPEECH DATA

This study used the TIMIT speech corpus. TIMIT consists of sentences read by 630 native speakers of eight major dialect regions of American English. It is labeled and segmented by hand in terms of 59 segments; i.e., the 50 phones listed in Table II, six labels for the closure parts of stop consonants (one for each stop; the stops listed in Table II only refer to the release parts of the stop consonants), and three labels for silence, which were collapsed onto one [sil] segment (see last row Table II). Of the 630 speakers in the corpus, 438 (70%) were male. For the analyses, TIMIT's standard test set was used (excluding the sa utterances). The test set used consists of 1344 utterances; 168 speakers each produced eight utterances from a set of 624 different utterances. The average number of boundaries per utterance is 36.5.

V. ANALYZING THE HYPOTHESIZED AND MISSED BOUNDARIES

The number of boundaries hypothesized by the MMC algorithm was similar to the number of boundaries in the transcriptions of the test set of TIMIT. The 1344 utterances of TIMIT's test set contain 45 514 boundaries, while the speech segmentation algorithm hypothesized 44 885 boundaries (resulting in an under-segmentation of 1.4%). Taking the TIMIT manual segmentations as ground truth, 30 926 (67.9%) of the hypothesized boundaries were correctly hypothesized, i.e., they appeared within a distance of 20 ms of the manually placed boundaries in TIMIT. Thus, 14 588 boundaries were missed. The algorithm also hypothesized 13 959 boundaries that do not coincide with segment boundaries in TIMIT; we refer to these as additionally hypothesized boundaries. In these cases, there is apparently a difference between clusters of frames inside the sliding window that is big enough to warrant hypothesizing a boundary, even though there is no segment boundary according to the TIMIT segmentation.

TABLE I. 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	+voice, −voice
Height	High, mid, low, nil, silence
Backness	Front, central, back, nil
Roundness	+round, −round, nil
Staticity	Static, dynamic

We carried out an in-depth analysis investigating when these additional boundaries occur and when boundaries are missed taking the TIMIT manual segmentations as ground truth. In previous analyses (Scharenborg *et al.*, 2007), we found that, like for automatic speech recognition systems, silence is problematic for our unsupervised speech segmentation algorithm. More specifically, the end of a silence tended to be hypothesized poorly due to problems with the endpointing algorithm. To avoid the endpointing problem, in the current study, we only investigated those boundaries that occurred at least 45 ms from the start or before the end of the reference TIMIT file. This resulted in 10 884 additionally hypothesized boundaries and 13 385 missed boundaries. Note that there are still silence frames, for instance, due to silences between words.

A. Set-up of the analyses

In order to be able to generalize over different segments, we characterized segments by "articulatory features" (AFs). AFs are the acoustic correlates of articulatory properties of speech sounds. We used the set of seven articulatory features shown in Table I. 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]). It might seem that the AF *staticity* will correlate almost perfectly with the hypothesized boundaries; however, this is not the case as *staticity* is related to the bigger manifestations of acoustic change, whereas our segmentation algorithm is also sensitive to smaller acoustic changes. [nil] is used when an AF is not applicable for a segment, for instance, consonants do not have a value for

TABLE II. Feature value specification of each TIMIT phone label in the test set. Note that the affricates /tʃ, dʒ/ are classified as fricatives since the largest parts of these segments are continuous.

Phone	Manner	Place	Voice	Height	Backness	Roundness	Staticity
æ	Vowel	Nil	+voice	Low	Front	−round	Static
ɛ	Vowel	Nil	+voice	Mid	Front	−round	Static
aʊ	Vowel	Nil	+voice	Low	Front	−round	Dynamic
eɪ	Vowel	Nil	+voice	Mid	Front	−round	Dynamic
aɪ	Vowel	Nil	+voice	Low	Front	−round	Dynamic
ɨ	Vowel	Nil	+voice	High	Front	−round	Static
i	Vowel	Nil	+voice	High	Front	−round	Dynamic
ɪ	Vowel	Nil	+voice	High	Front	−round	Static
ə	Vowel	Nil	+voice	Mid	Central	−round	Static
ʌ	Vowel	Nil	+voice	Mid	Central	−round	Static
ɔ	Vowel	Nil	+voice	Low	Back	+round	Static
oʊ	Vowel	Nil	+voice	Mid	Back	+round	Dynamic
oɪ	Vowel	Nil	+voice	Low	Back	+round	Dynamic
ʊ	Vowel	Nil	+voice	High	Back	+round	Static
u	Vowel	Nil	+voice	High	Back	+round	Dynamic
ɑ	Vowel	Nil	+voice	Low	Back	+round	Static
ʉ	Vowel	Nil	+voice	High	Back	+round	Dynamic
l	Approximant	Alveolar	+voice	Nil	Nil	Nil	Dynamic
ɭ	Approximant	Alveolar	+voice	Nil	Nil	Nil	Dynamic
w	Approximant	Velar	+voice	Nil	Nil	Nil	Dynamic
j	Approximant	Velar	+voice	Nil	Nil	Nil	Dynamic
p	Stop	Bilabial	−voice	Nil	Nil	Nil	Dynamic
b	Stop	Bilabial	+voice	Nil	Nil	Nil	Dynamic
t	Stop	Alveolar	−voice	Nil	Nil	Nil	Dynamic
d	Stop	Alveolar	+voice	Nil	Nil	Nil	Dynamic
k	Stop	Velar	−voice	Nil	Nil	Nil	Dynamic
g	Stop	Velar	+voice	Nil	Nil	Nil	Dynamic
ɾ	Stop	Alveolar	+voice	Nil	Nil	Nil	Dynamic
f	Fricative	Labiodental	−voice	Nil	Nil	Nil	Static
v	Fricative	Labiodental	+voice	Nil	Nil	Nil	Static
θ	Fricative	Dental	−voice	Nil	Nil	Nil	Static
ð	Fricative	Dental	+voice	Nil	Nil	Nil	Dynamic
s	Fricative	Alveolar	−voice	Nil	Nil	Nil	Static
ʃ	Fricative	Alveolar	−voice	Nil	Nil	Nil	Static
z	Fricative	Alveolar	+voice	Nil	Nil	Nil	Static
ʒ	Fricative	Alveolar	+voice	Nil	Nil	Nil	Static
h	Fricative	Velar	−voice	Nil	Nil	Nil	Static
ɦ	Fricative	Velar	+voice	Nil	Nil	Nil	Static
tʃ	Fricative	Alveolar	−voice	Nil	Nil	Nil	Dynamic
dʒ	Fricative	Alveolar	+voice	Nil	Nil	Nil	Dynamic
m	Nasal	Bilabial	+voice	Nil	Nil	Nil	Static
n	Nasal	Alveolar	+voice	Nil	Nil	Nil	Static
ŋ	Nasal	Velar	+voice	Nil	Nil	Nil	Static
ɱ	Nasal	Bilabial	+voice	Nil	Nil	Nil	Dynamic
ɳ	Nasal	Alveolar	+voice	Nil	Nil	Nil	Dynamic
ŋ	Nasal	Velar	+voice	Nil	Nil	Nil	Dynamic
ɾ̃	Nasal	Alveolar	+voice	Nil	Nil	Nil	Static
ɚ	Retroflex	Alveolar	+voice	Nil	Nil	Nil	Dynamic
ɹ	Retroflex	Alveolar	+voice	Nil	Nil	Nil	Dynamic
ɝ	Retroflex	Nil	+voice	Nil	Nil	Nil	Dynamic
Sil	Sil	Sil	−voice	Sil	Nil	Nil	Static

front-back (as their place of articulation is defined with the AF *place*) or *roundness*. [sil] is used for silent frames. Note that for *voice*, silence is marked as [-voiced], and for *staticity* as [static]. A final observation: in TIMIT, the silence (i.e., the closure) and release (i.e., the burst) parts of stops have been annotated separately, but in our study the transcription of a sequence of a silence part followed by a release part is changed to represent a single segment. Table II presents an overview of the feature value specification of each of the phone labels in the TIMIT set.

In the first series of analyses (Sec. V B), we investigated the presence of additional boundaries. In the second series of

analyses (Sec. V C), we investigated the segment contexts in which boundaries were missed. Section V D summarizes the most salient results of the analyses. The analyses were carried out using generalized linear mixed-effect models, thus containing both fixed and random predictors, using the logit link function. The fixed predictors are the AFs we defined (see Table I) and the duration of a segment (see below). The random predictors are described below. The parameters of the generalized linear models are set using maximum likelihood estimation. We used contrast coding.[1] A generalized model, with the logit link function, has the form

$$\text{logit } p = c\beta_I \text{AF}_I + \beta_2 \text{AF}_2 + \beta_3 \text{AF}_3 + \cdots + \beta_N \text{ duration},$$

where logit p represents $\log [p(1-p)]$. In our case, p is the probability that a boundary is inserted inside a segment or the probability of missing a boundary (logit p is the "dependent variable"). The constant c is the intercept. The different β's (Chatterjee *et al.*, 2000) represent the relevance (effect size) of the different AFs and duration for the estimation of the logit p: a larger absolute β corresponds to a larger effect of the corresponding predictor.

In the following analyses, only statistically significant (calculated using F-tests) effects are part of the final statistical model and reported. In addition, we report the absolute estimated values of the different β's, with an explanation of whether the likelihood of an additional or missed boundary increases or decreases with the associated feature.

B. Additionally hypothesized boundaries

We investigated whether the presence of an additional boundary within a segment can be predicted based on its context. As indicated above, the algorithm hypothesized 10 884 additional boundaries. When a boundary falls within a segment of which the initial boundary or final boundary is missing, it might be the case that the additional boundary is, in fact, a misallocated initial/final boundary instead of a "true" additionally hypothesized boundary. We, therefore, only investigated those 19 033 segments for which the initial and final boundaries were correctly hypothesized. Of these 19 033 segments, 4079 segments had an additional boundary; these were compared to the 14 954 segments that did not have an additional boundary. Thus, 37.5% (4079 of 10 884) of the additional boundaries hypothesized by the algorithm were true additional boundaries. The number of segments with additional boundaries thus is high enough to be able to detect patterns in the errors made by the unsupervised segmentation algorithm.

For each target segment, as well as its preceding and succeeding segment, the phone label in the TIMIT transcription was determined. Additionally, the segment was rewritten in terms of its AF values (see Table II, for an overview). The duration and the AF values of the target segment, in addition to all two-way interactions, were tested as fixed predictors. Segment durations were calculated from the hand-segmented TIMIT data. The mean segment duration over all analyzed segments was 80.8 ms. As random predictors, speaker identity and the phone identities of the target segment itself and of the preceding and succeeding segments were tested.

We ensured none of the variables correlated with one another in the analyses. In case a variable correlated with another, we either removed the variable from the analysis or reorganized the levels of the variable, as described below. However, for duration we followed a different path. The duration of a segment turned out to correlate with some of the AF values of the segment itself. We orthogonalized the duration of a segment with these AF values. We let the duration be predicted by the AF values with which it correlates, and used the residuals of this model as the "duration" predictor in the analyses. This residual duration thus is the duration that cannot be predicted on the basis of the AF values. This procedure ensures that if we find an effect for duration, it indeed can be attributed to duration and not to the AF values. The residual duration is calculated for every analysis separately.

Since only *manner* can be meaningfully specified for all segments, we first investigated whether the presence of an additional boundary within a segment can be predicted by the *manners* of articulation of the phone sequence. Of the random predictors, speaker identity as well as the phone identities of the target and the preceding segment appeared to contribute to the explanation of the variation ($p<0.05$). Furthermore, we observed robust effects of the residual duration of the segment [$\beta=0.0220$, $F(1,19025)=1399.883$, $p<0.0001$]: additional boundaries are more likely in longer segments. Second, we observed robust effects of the manner of articulation of the segment under analysis [$F(6,19025)=11.646$, $p<0.0001$].

The primary reason that longer segments are more likely to have additional boundaries is because of the applied segmentation algorithm. MMC analyzes a fixed number of frames equivalent to 90 ms of speech. If a segment is longer than 90 ms, the analysis window will contain frames coming from one phone only. MMC will always hypothesize a boundary, even if the acoustic change within the segment is very small as the MMC algorithm is sensitive to the tiniest of changes in the MFCC feature space. If the acoustic change is big enough, it is picked up by the peak detector, or if the MMC algorithm places the hypothesized boundaries consistently over time, the boundary is picked up by the mask detection method (see Sec. III). Both mechanisms result in the segment boundary being hypothesized. In the case of shorter segments whose durations match the size of the analysis window, the boundary hypothesized by MMC is more likely to indeed be a segment boundary because the frames in the analysis window will be from two segments, thus resulting in fewer additional boundaries.

Further statistical analyses allowed us to establish four groups of manner values with the following general trends. As is shown in Fig. 3, stops appeared to have most additional boundaries (32.8% of the stops had an additional boundary), followed by fricatives (25.2%) and silence (23.6%), followed by vowels (20.7%), while retroflexes (14.0%), nasals (7.8%), and approximants (3.4%) appeared to have the smallest number of additional boundaries. It is not surprising that stops get more additional boundaries than any other type of segments. The algorithm is designed to group together frames that are similar. Since stops consist of two distinct parts (remember that the closure and release part of stops are labeled as one

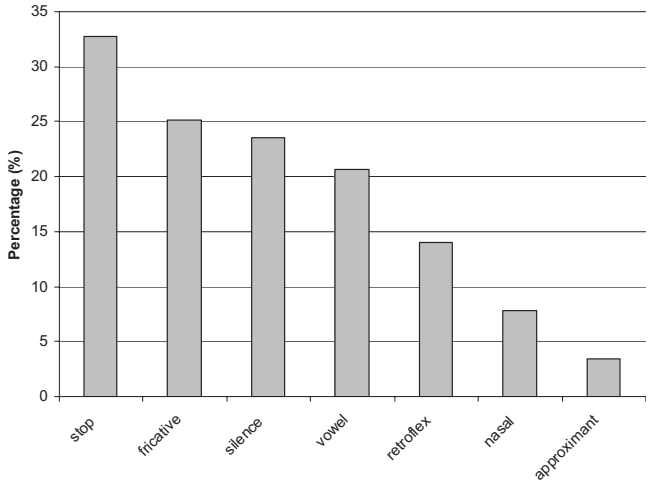

FIG. 3. The percentage of segments containing an additional boundary per *manner* class.

segment—contrary to the standard TIMIT labeling), the algorithm tends to divide the stop into two separate segments resulting in the hypothesis of additional boundaries. Fricatives might show more variation than sonorant consonants resulting in more additional boundaries for fricatives. In case of silence, as the MMC algorithm is sensitive to the tiniest of acoustic change, it is likely that it picks up small background noises, for instance, due to speaker noises or the microphone, resulting in additional boundaries.

The acoustic realizations of vowels, retroflexes, nasals, and approximants are very similar due to the lack of a stricture of the vocal tract sufficient to cause audible turbulence during the production of these segments. It might therefore be surprising that vowels behave differently from retroflexes, nasals, and approximants in that vowels get more additional boundaries. This difference can most likely be attributed to the fact that 29.0% of the [vowel] segments are diphthongs; during the realization 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 vowel into two separate segments. Retroflexes, nasals, and approximants, on the other hand, are more or less static sounds, and this staticity results in fewer additionally hypothesized boundaries.

In order to test the role of the other AFs in the hypothesizing of additional boundaries, we analyzed obstruents, nasal consonants, and vowels separately, and investigated which of their characteristics predict the presence of an additional boundary. For the obstruents, *voice*, *manner* (thus either [stop] or [fricative]), *staticity*, and *place* of articulation are meaningful AFs. The predictors *manner* and *staticity* are correlated, and in order to avoid collinearity, *staticity* was taken out as a predictor. Furthermore, *place* and *manner* of articulation are correlated. To remove this correlation, we relabeled the values of *place* into three levels: [front], [middle], and [back], see Table III for an overview.

We orthogonalized duration with the remaining predictors. The resulting model showed that of the random predictors only the identities of the target and following segments contributed to the explanation of the variation ($ps < 0.0001$). Furthermore, we found a main effect for residual duration [$\beta = 0.0197$, $F(1, 19025) = 626.105$, $p = 0$] and an interaction for residual duration and *manner*, while the main effect of *manner* was not significant: additional boundaries are more often hypothesized in longer stops than in fricatives [$\beta = 0.0155$, $F(1, 19025) = 50.374$, $p < 0.0001$]. Residual duration also interacted with *voice* [$\beta = 0.0106$, $F(1, 19025) = 14.029$, $p < 0.001$], which also showed a main effect [$\beta = 1.3723$, $F(1, 19025) = 30.187$, $p < 0.0001$]: a [+voice] segment has a smaller likelihood for getting additional boundaries than a [−voice] segment; however, this difference is attenuated with increasing residual duration of the segment.

As explained above, stops are more likely to get additional boundaries than fricatives. In longer stops, it is to be expected that the additional boundary is placed further away from the end boundary of the stop, and this boundary is therefore maintained instead of being "smoothed" away (see Sec. III), as may be expected for shorter stops. This will result in more additional boundaries especially for longer stops than for fricatives or shorter stops.

We expected the acoustic change occurring in [+voice] segments to be less than the acoustic change occurring in [−voice] segments for two reasons. First, the closure part of voiced stops may contain voicing, resulting in less acoustic change from the closure to the burst compared to going from a silent closure to a burst in voiceless stops. Second, the burst in voiced stops are not as pronounced as the bursts of voiceless stops, which also results in less acoustic change for the voiced segments. This difference between voiced and voiceless stops attenuates for longer segments, because of the way MMC performs segmentation. As explained above, if a segment is longer than 90 ms, the analysis window will contain frames coming from one phone only, resulting in

TABLE III. Overview of the relabelling of the place AF values for the obstruent analysis.

		#fricatives		#stops	
AF value		Old	New	Old	New
[front]	[bilabial]	0	1284	990	990
	[labiodental]	940		0	
	[dental]	344		0	
[middle]	[alveolar]	2507	2507	1084	1084
[back]	[velar]	204	204	878	878

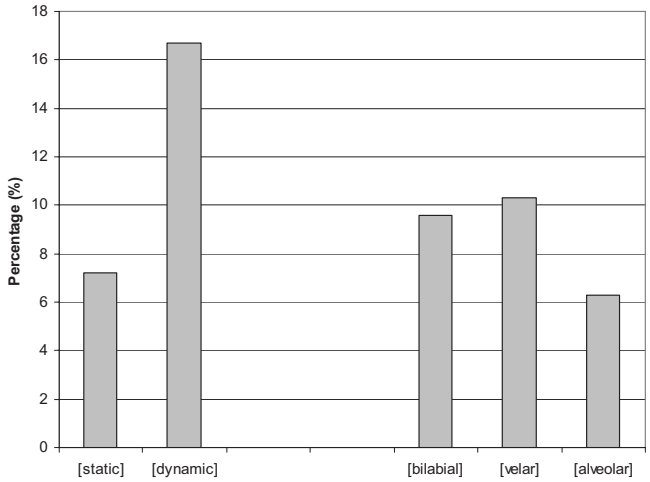

FIG. 4. Results for the analysis of the [nasal] segments: The percentage of segments containing an additional boundary per *staticity* feature value (left) and per *place* feature value (right).

more additional boundaries irrespective of the type of segment. This attenuating effect of duration was also found in other analyses. As it is explained here and above, we will not come back to it below.

As *voice* and *manner* are not meaningful predictors for [nasal] segments, we only studied the role of *staticity* and *place* of articulation in addition to the random predictors speaker identity and the phone identities of the target, preceding, and succeeding segment, for [nasal] segments. In TIMIT, four static and three dynamic nasals occur. The [static] nasals are /m, n, ŋ, ɾ̃/ (/ɾ̃/ is the nasal flap in American English as in wintergreen); the [dynamic] nasals are syllabic /m̩, n̩, ŋ̩/. Duration was again orthogonalized. *Place* only consisted of three AF values, i.e., [bilabial], [alveolar], and [velar], as [labiodental] and [dental] nasals do not occur in English.

Of the random predictors, only phone identity of the succeeding segment appeared to contribute to the explanation of the variation. We found main effects for the residual duration of the segment: additional boundaries were more likely for increasing segment duration [$\beta=0.0410$, $F(1,19025)=149.264$, $p<0.0001$]. *Staticity* also showed a main effect [$\beta=1.2210$, $F(1,19025)=24.174$, $p<0.0001$]: additional boundaries were less often hypothesized in [static] segments than in [dynamic] segments (7.2% of the [static] nasals had additional boundaries compared to 16.7% of the [dynamic] nasals; see also Fig. 4). Finally, *place* of articulation showed a main effect [$F(2,19025)=4.9334$, $p=0.01$].

Like the effect of duration (see the manner analysis for an explanation), the effect of *staticity* is as expected. *Staticity* indicates whether an acoustic change occurs in the segment or not, and as explained above, dynamic change is the basis for hypothesizing boundaries (as is for most unsupervised speech segmentation algorithms). It is thus to be expected that [dynamic] segments more often have an additional boundary than [static] segments.

Further statistical analyses showed that additional boundaries were more likely for [bilabial] (9.6%) and [velar] (10.3%) nasals than for [alveolar] nasals (6.3%), which is also shown in Fig. 4. The movements of the articulators are comparatively bigger when producing a constriction at the lips (for bilabial nasals) or in the back of the oral cavity (for velar nasals) compared to producing a constriction close to the alveolar ridge (for alveolar nasals). The transitions are thus comparatively greater for [bilabial] and [velar] nasals, which implies more acoustic change and thus more additionally hypothesized boundaries for [bilabial] and [velar] nasals.

Vowels differ in their specification for *height, backness, roundness*, and *staticity*. With respect to the AF values for staticity, all diphthong vowels are marked as [dynamic], while all monothongs are marked as [static]. *Backness* correlated with *roundness*, *height*, and *staticity* and was thus removed from the analysis. Duration was then orthogonalized based the remaining three predictors.

Two random predictors appeared to contribute: speaker identity and the phone identity of the target segment ($ps<0.05$). We again found a main effect for residual duration [$\beta=0.0153$, $F(1,19025)=364.728$, $p<0.0001$]: additional boundaries were more likely with increasing duration. The second predictor that showed a main effect was *staticity* [$\beta=0.8915$, $F(1,19025)=38.122$, $p<0.0001$]: as before, the likelihood for additional boundaries was smaller for [static] segments than for [dynamic] segments; however, this difference attenuated with increasing residual segment duration [$\beta=0.0067$, $F(1,19025)=19.334$, $p<0.0001$]. Third, *height* [$F(2,19025)=40.446$, $p<0.0001$] showed a main effect: additional boundaries were more likely to be hypothesized in [low] vowels than in [high] and [mid] vowels; however, also this difference disappears for increasing residual duration of the segment [$\beta=0.0072$, $F(2,19025)=6.377$, $p<0.005$]. Finally, *roundness* [$\beta=0.3490$, $F(1,19025)=10.883$, $p<0.0001$] showed a main effect: additional boundaries were less often positioned in [−round] vowels than in [+round] vowels ([−round]: 19.4% vs [+round]: 29.1%).

The higher number of additional boundaries for [dynamic] and for longer segments is again in agreement with the results presented above. During the production of a [low] vowel, the mouth is more open than during the production of a [high] or [mid] vowel, while for the production of the constriction of preceding and following consonants, the mouth needs to be fairly, or even entirely, closed. As a consequence, the formant transitions are comparatively greater in [low] vowels, which implies more acoustic change and thus more additionally hypothesized boundaries than in [high] and [mid] vowels. A similar explanation holds for the difference in the likelihood of additional boundaries for [+round] and [−round] segments. In order to produce a round vowel, there is more lip movement involved compared to the production of an unround vowel, which results in more acoustic change, and thus an increase in additional boundaries for round vowels.

C. Missed boundaries

We subsequently investigated whether a missing boundary can be predicted based on its segment context. For these analyses, we followed the general procedure of the analysis of the additional boundaries. As indicated above, the algo-

rithm missed 13 385 boundaries. To ensure a boundary was indeed missed and not merely shifted in time, we restricted our analyses to the 7733 missed boundaries that were not preceded or followed by additional boundaries. These were compared to 18 105 boundaries that were not missed and also were not preceded or followed by additional boundaries. Thus, 57.8% (7733 of 13 385) of the missed boundaries were true missed boundaries. Again, the number of missed segments is high enough to be able to detect patterns in the errors made by the unsupervised segmentation algorithm.

Acoustic change leads to the hypothesizing of boundaries. So, if the frames on either side of the boundary are similar, boundaries are more likely to be missed. We therefore expect that only the agreement in *manner* class is a meaningful predictor. This is easily illustrated with a few examples. Imagine a segment [f] followed by [æ]. [f] and [æ] are both [static] segments (see Table II). Even though there is an agreement in *staticity*, the acoustic change between the two segments is rather big, since this is a transition from a fricative to a vowel. Likewise, imagine the transition from an [l] to a [d] sound. Even though there is agreement in *voice* and *place*, the acoustic change occurring when going from the first to the second segment will be quite big again, as this is a transition from an approximant to a stop. We hypothesize that the frames on either side of a boundary are similar when the *manner* of articulation is similar, and thus that boundaries are more likely to be missed when there is an agreement in *manner* AF value for the two segments on either side of the boundary. We tested this in our analysis.

For each (missed or present) boundary, we determined the phone label of the preceding and following segment from the TIMIT transcription and their manner AF values (following Table II). The durations of the segments turned out to correlate with their manners of articulation; therefore, the duration of the preceding segment was orthogonalized with the *manner* of the preceding segment; likewise, the duration of the succeeding segment was orthogonalized with the *manner* of the succeeding segment.

We created a new variable indicating whether the two *manner* AF values of the surrounding segments were similar or dissimilar. We grouped together approximants, vowels, retroflexes, and nasals since for all these sounds the constriction of the vocal tract is minimal, this in contrast to fricatives and stops where there is a clear closure or audible turbulence. [silence] segments were grouped together with fricatives and stops, since stops also have silent portions in them.

The residual durations, the recoded manner AF values of the preceding and succeeding segments, and the two-way interactions were tested as fixed predictors. As crossed random factors, the phone identities of the two segments, as well as the speaker identity were tested. Only those predictors that proved to be significant were kept in the model.

The resulting model showed that of the random predictors only identities of the preceding and succeeding segments contribute to the explanation of the variation. We found a main effect for the agreement in *manner* between the two segments [$\beta=1.154$, $F(1,25832)=853.697$, $p<0.0001$]. As expected, boundaries are more likely to be missed when the segments on either side of the boundary have the same *manner* class. The acoustic changes occurring when going from a fricative, silence, or stop to a retroflex, nasal, vowel, or approximant, or vice versa, are far greater than when going from a segment from one class to a segment from the same class. Furthermore, we found a main effect for the residual duration of the preceding segment [$\beta=0.0279$, $F(1,25832)=681.684$, $p<0.0001$]: the likelihood of missing a boundary decreases with increasing duration of the preceding segment; however, this is less so when both segments have the same *manner* class [$\beta=0.0190$, $F(1,25832)=235.010$, $p<0.0001$]. The residual duration of the succeeding segment also showed a main effect [$\beta=0.0102$, $F(1,25832)=218.029$, $p<0.0001$]: again, an increasing duration reduces the likelihood of missing a boundary. The effect of the duration of the succeeding segment attenuates with increasing duration of the preceding segment, and vice versa [$\beta=0.00004$, $F(1,25832)=11.768$, $p<0.05$].

As explained above, the MMC analysis window is exactly 90 ms long, and MMC only hypothesizes one boundary per analysis window. Thus, in case segments are much shorter than 90 ms, especially when there are multiple segments embedded in the 90 ms analysis window, MMC will miss some of the boundaries between the segments. With increasing segment duration, the analysis window will no longer contain more than two segments, but two or only one segment; this thus results in a smaller likelihood of missing boundaries. Nevertheless, when the segments on either side of the boundary are similar, there is little acoustic change and thus the likelihood of missing that boundary increases compared to when the segments on either side are dissimilar, as the boundary detection method will be less likely to detect the acoustic change.

D. Summary

As is clear from the above analyses of the additional boundaries, the duration of the segments plays a major role in predicting the presence of an additional boundary. In all analyses, duration showed a main effect. If segments are longer, it is more likely that the frames in the 90 ms analysis window all belong to the same segment. MMC will always hypothesize a boundary, even if the acoustic change within the segment is very small as the MMC algorithm is sensitive to the tiniest of changes in MFCC feature space. If the acoustic change is big enough, it is picked up by the peak detector, or if the boundaries were consistently placed over time by the MMC algorithm it is picked up by the mask detection method (see Sec. III), resulting in an additional boundary. The second major finding is that stops get more additional boundaries than any other type of segment. MMC is designed to group together frames that are similar. Since stops typically consist of two distinct parts, whereas our transcription of the [stop] segment puts both parts into one segment, the algorithm often divides the stop into two separate segments (so the boundary detection method is able to detect the boundary), resulting in the hypothesis of additional boundaries. This effect of *manner* is attenuated in longer obstruents, as all long segments are equally prone to additional boundaries due to the characteristics of the applied segmen-

tation algorithm. The fourth major finding is the effect of *staticity*. *Staticity* indicates whether an acoustic change occurs in the segment or not. As explained above, dynamic change is the basis for hypothesizing boundaries. It was thus to be expected that [dynamic] segments more often have an additional boundary than [static] segments.

The analyses of missed boundaries showed that duration also plays a major role in predicting the presence or absence of a boundary: the likelihood of hypothesizing a boundary increases with increasing duration of the preceding and succeeding segment. As explained above, MMC only hypothesizes one boundary per analysis window. In case segments are much shorter than 90 ms, especially when there are multiple segments in the analysis window, MMC will miss some of the boundaries. With increasing segment duration, the analysis window will contain fewer segments thus reducing the likelihood of missing boundaries. The second interesting finding is that boundaries are more likely to be missed when the segments on either side of the boundary have the same *manner* class, which is to be expected since when frames on either side of the boundary are similar, the acoustic change is smaller, thus increasing the likelihood of missing the boundary.

VI. GENERAL DISCUSSION

Despite using different algorithmic implementations, most unsupervised speech segmentation methods achieve similar performance in terms of percentage correct boundary detection (at similar under- or over-segmentation rates; see Sec. II). Nevertheless, unsupervised speech segmentation algorithms are not able to perfectly reproduce manually obtained reference transcriptions. We are interested in trying to unearth the fundamental problems for unsupervised automatic speech segmentation algorithms. To that end, we compared a phone segmentation obtained using only the acoustic information present in the signal with a reference segmentation created by human transcribers, and analyzed the boundaries that were additionally hypothesized and those that were missed.

The comparison of the different unsupervised speech segmentation algorithms and their performances in Sec. II showed that more sophisticated automatic segmentation algorithm will likely result in an improved segmentation algorithm. However, in such more sophisticated systems, it is more difficult to tease apart the effects of different parts, e.g., related to the heuristics, assumptions, or signal processing, of the segmentation algorithm; furthermore, the results become specific to that segmentation algorithm. The criterion that often underlies the decision process for the hypothesis of boundaries in unsupervised automatic speech segmentation algorithms is acoustic change. This let us to the question how good an indicator of a segment boundary acoustic change is. To answer this question, we chose to analyze the results of an automatic segmentation algorithm that only uses acoustic change as the criterion for hypothesizing a segment boundary, and that produced enough errors that made it possible to detect patterns in the errors. This enterprise indicated where acoustic change is indeed a correct indicator of a segment boundary, and where it is not, thus revealing fundamental problems for automatic speech segmentation algorithms.

It is important to note though that the criterion of acoustic change can be applied in several ways with different underlying assumptions of which ours is just one. As a consequence, the results and thus the analyses presented in this paper are still somewhat tied to the speech segmentation algorithm that was used. Nevertheless, we believe that the results found are directly related to the acoustic change criterion and will hold for most other speech segmentation algorithms based on this criterion.

The analyses showed that acoustic change indeed is a fairly good indicator of segment boundaries: 67.9% of the boundaries hypothesized by the MMC algorithm coincide with segment boundaries when there is no over-segmentation. The analyses showed that the MMC algorithm is sensitive to very subtle changes within a segment: it was able to pick up acoustic changes as great as the transition from the closure to the burst of a plosive, but also as subtle as the formant transitions in [low] vowels.

So, why are some boundaries erroneously inserted or missed? The analyses showed that the errors made by the unsupervised speech segmentation algorithm can be split into three groups. Here, we will address these three groups of errors made by the segmentation algorithm. Below, we will give suggestions as to how we believe these errors can (partially) be dealt with in order to improve unsupervised speech segmentation. First of all, the MMC algorithm has problems related to segment duration. Second, the MMC algorithm has problems dealing with adjacent segments that are similar; boundaries are likely to be missed if both segments have the same or a similar manner class, e.g., a nasal followed by a vowel. We should note, however, that this is also problematic for human transcribers. Third, the algorithm has no means of dealing with inherently dynamic phones; the algorithm, for instance, often hypothesizes boundaries between the closure and the burst in stops resulting in two separate segments instead of one stop segment. We believe that most, if not all, of these problems will occur in most current algorithms of unsupervised speech segmentation that use acoustic (rate of) change as the only, or the most important, means to segment speech. In order to develop unsupervised speech segmentation algorithms that are able to create segmentations that are closer to those created by human transcribers, these three main problems need to be dealt with.

A partial solution to the issue of segment duration is related to the way MMC performs speech segmentation. In the analysis, a fairly long window of 90 ms was used which means that we may miss some boundaries in a series of short segments, as the MMC algorithm can only hypothesize one boundary per window. However, a short(er) analysis window is more likely to contain frames belonging to one segment only, resulting in more additional boundaries. One way to improve the algorithm's performance for short segments is to use a multi-class MMC (Zhao *et al.*, 2008), instead of the two-class MMC we used so far; this will enable the detection of several clusters within one analysis window. Another way of improving the algorithm is to use an analysis window that changes in size dynamically or performs the analysis using

multiple window sizes; this could also help reduce the number of additional boundaries in longer segments. This will also make it possible to deal with segments that inherently differ in duration, e.g., consonants are in principle shorter in duration than vowel. A third method is an approach or model that is able to keep track of the duration (as a function of speech rate) and knows when the right conditions are met to hypothesize a boundary. One way of doing this is to make the threshold δ of the Euclidean distance method dependent on the distance to the last hypothesized boundary. This might, however, result in an over-segmentation. Future research will shed light on these possible solutions.

One way of dealing with boundaries that are missed between adjacent segments where acoustic change between the segments is small is to allow the MMC algorithm to oversegment and then apply rules to filter away the additional hypotheses. An example of such a rule might be the removal of the additional boundary that is hypothesized in stops. For instance, if a segment preceding a boundary contains silence or murmur and the segment following the boundary contains a burst, the boundary between the two segments should be removed. These rules may be formulated based on the statistical regularities in large datasets. Possibly, some of these rules may need to be language-dependent. This line of reasoning suggests that we need to know what the sound is in order to be able to segment the speech more accurately.

Unsupervised speech segmentation algorithms are necessarily bottom-up approaches, since they have no prior knowledge of the material (other than the acoustic signal). They hypothesize boundaries without knowledge of possible acoustic phenomena related to the transition from one segment to the next. They have no knowledge about the number of segments there are in the speech material they are supposed to segment, nor about the types of labels of the segments. This thus implies that rules based on durational information, information about the (dis)similarity of adjacent segments, or inherently dynamic phones cannot be obtained in a bottom-up fashion. So perhaps unsupervised speech segmentation can only be improved by including so-called top-down information, i.e., information about segment labels, as is used by supervised speech segmentation algorithms. However, this would be undesirable for reasons listed in Sec. I.

In keeping with the wish to build a system that is flexible and language-independent, we propose a system that, rather than taking the fully supervised approach of training a predefined set of phone models, automatically clusters the hypothesized segments that are derived from acoustic change into broad classes, such as voiced or voiceless, classes for different durations, and/or classes associated with some of the other AF values based on the (dis)similarities in the acoustic signal, and then reconsider the hypotheses accordingly. We refer to this type of information as automatically derived top-down information. This approach would yield a multi-stage system, consisting of a mix of bottom-up and automatically derived top-down information used for the task of speech segmentation. In such a system, the first stage would consist of the original bottom-up unsupervised speech segmentation algorithm; its output would subsequently be smoothed by a model based on the labels of the broad classes and the acoustic signal in a separate step.

To summarize, the analyses showed that acoustic change indeed is a fairly good indicator of segment boundaries: over two-thirds of the boundaries hypothesized by the MMC algorithm coincide with segment boundaries when there is no over-segmentation. The remaining errors highlighted the fundamental problems for unsupervised automatic speech segmentation methods; these are related to segment duration, sequences of similar segments, and inherently dynamic phones. In order to improve unsupervised automatic speech segmentation, we suggest current one-stage bottom-up segmentation methods to be expanded into two-stage segmentation methods that are able to use top-down information based on automatically derived broad classes and rules. In this way unsupervised methods can be improved while remaining flexible and language-independent. Note, however, that some rules might be language-dependent. Obviously, when including these rules, the resulting two-stage segmentation method will no longer be language-independent.

To conclude, it is difficult to hypothesize what would happen to our analyses when they would have been carried out with a more sophisticated speech segmentation algorithm. We therefore would like to encourage future papers on new algorithms of unsupervised automatic speech segmentation to also include an analysis of the errors, along the lines of our analyses of additionally hypothesized and missed boundaries, so it will become clear where the improvement of the more sophisticated segmentation algorithm originates.

ACKNOWLEDGMENTS

The results presented in this article supersede earlier presented results published in the Proceedings of ICASSP 2007, Honolulu, HI, and the Proceedings of Interspeech 2007, Antwerp, Belgium. This research was supported by a Veni-grant from the Netherlands Organisation for Scientific Research (NWO) to Odette Scharenborg and by a EURYI-award from the European Science Foundation to Mirjam Ernestus.

[1]One segment or combination of AF values is used as the "Intercept," i.e., the default, to which all other segments or combinations of AF values are compared.

Aversano, G., Esposito, A., Esposito, A., and Marinaro, M. (**2001**). "A new text-independent method for phoneme segmentation," in Proceedings of the 44th IEEE Midwest Symposium on Circuits and Systems, Fairborn, OH, Vol. **2**, pp. 516–519.

Bridle, J., and Sedgwick, N. (**1977**). "A method for segmenting acoustic patterns, with applications to automatic speech recognition," in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 656–659.

Brugnara, F., Falavigna, D., and Omologo, M. (**1993**). "Automatic segmentation and labeling of speech based on hidden Markov models," Speech Commun. **12**, 357–370.

Burges, C. J. C. (**1998**). "A tutorial on support vector machines for pattern recognition," Data Min. Knowl. Discov. **2**, 1–47.

Chatterjee, S., Hadi, A. S., and Price, B. (**2000**). *Regression Analysis by Example* (Wiley, New York).

Cucchiarini, C. (**1993**). "Phonetic transcription: A methodological and em-

pirical study," Ph.D. thesis, University of Nijmegen, Nijmegen, The Netherlands.
Dusan, S., and Rabiner, L. (**2006**). "On the relation between maximum spectral transition positions and phone boundaries," in Proceedings of Interspeech, Pittsburgh, PA.
Garofolo, J. S. (**1988**). "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology (NIS), Gaithersburgh, MD.
Kim, Y.-J., and Conkie, A. (**2002**). "Automatic segmentation combining an HMM-based approach and spectral boundary correction," in Proceedings of International Conference on Spoken Language Processing, Denver, CO, pp. 145–148.
Kuperman, V., Pluymaekers, M., Ernestus, M., and Baayen, R. H. (**2007**). "Morphological predictability and acoustic duration of interfixes in Dutch compounds," J. Acoust. Soc. Am. **121**, 2261–2271.
Pellom, B. L., and Hansen, J. H. L. (**1998**). "Automatic segmentation of speech recorded in unknown noisy channel characteristics," Speech Commun. **25**, 97–116.
Pereiro Estevan, Y., Wan, V., and Scharenborg, O. (**2007**). "Finding maximum margin segments in speech," in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Honolulu, HI.
Qiao, Y., Shimomura, N., and Minematsu, N. (**2008**). "Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons," in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Las Vegas, NV.
Scharenborg, O., Ernestus, M., and Wan, V. (**2007**). "Segmentation of speech: Child's play?" in Proceedings of Interspeech, Antwerp, Belgium, pp. 1953–1956.
Sharma, M., and Mammone, R. (**1996**). ""Blind" speech segmentation: Automatic segmentation of speech without linguistic knowledge," in Proceedings of International Conference on Spoken Language Processing, Philadelpia, PA, pp. 1237–1240.
Wesenick, M.-B., and Kipp, A. (**1996**). "Estimating the quality of phonetic transcriptions and segmentations of speech signals," in Proceedings of International Conference on Spoken Language Processing, Philadelpia, PA, pp. 129–132.
Xu, L., Neufeld, J., Larson, B., and Schuurmans, D. (**2004**). "Maximum margin clustering," in Proceedings of NIPS, Vancouver, Canada.
Zhao, B., Wang, F., and Zhang, C. (**2008**). "Efficient multiclass maximum margin clustering," in Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, pp. 1248–1255.