Using cepstral coefficients for Inhalation pause detection in spontaneous speech

Sjöström, Anders; Frid, Johan; Horne, Merle

Published in:
Proceedings of SPECOM 2005

2005

Using cepstral coefficients for inhalation pause detection in spontaneous speech

Anders Johansson, Johan Frid, Merle Horne

Department of Linguistics and Phonetics
Lund University, Lund, Sweden
anders.johansson@ling.lu.se, johan.frid@ling.lu.se, merle.horne@ling.lu.se

Abstract

A method for recognizing inhalations in spontaneous speech is presented. It is similar to the template matching technique; a distance measure is calculated between a reference sound and an equally long portion of the same sound being tracked. A feature representation consisting of the standard Mel Frequency Cepstral Cofficients (MFCC), obtained by performing a Discrete Cosine Transform of the mel-scaled filterbank spectrum is used. MFCCs are calculated for every 5 ms. The comparison is then done by computing the euclidean distance between the cepstral coefficents of each frame of the two sounds. A low distance value means that the two compared inhalations are likely to be similar. The method can detect inhalations in both male and female spontaneous speech. The method is most suited for signals with low noise and high average intensity (studio recordings) but can also be used on noisier recordings with lower average intensity, albeit with poorer results.

1. Introduction

During our recent investigations into the prosodic phrasing of speech fragments following hesitations, it has been observed that they seem to be grouped into units that have a relatively constant duration [5]. This idea finds support in memory research. For example, Baddeley [1] has claimed that the part of working memory where speech processing takes place ('inner speech') has a time limit of around 2 seconds.

Discussions of timing constraints in linguistic research can be found in studies on speech rhythm; for example, Fant and Kruckenberg [3] have focussed on the duration of inter-stress stretches of read speech. Empirical investigations on timing restrictions on larger speech production chunks/information units has not, to our knowledge, been the object of detailed investigation. Sigurd [7], however, assumes speech chunks of one to two seconds' length in his message-to-speech model. We would like to build on these ideas and to hypothesize that speech planning units are between 2-2.5 seconds long. We further hypothesize that they can contain internal silent and/or filled pauses, but not pauses containing inhalations.

The 2-2.5 second production units that we are envisaging can be thought to correspond to the ouput of the linguistic Formulator in Levelt's [6] model of speech production. In this model, one can expect that pauses internal to production units can be related to e.g. lexical access time. (It should be noted, however, that Levelt does not assume any timing restriction on speech coding in his model). Our evidence for this assumed timing restriction on speech production comes mainly from observations of prosodic phenomena (accentual and pausal) associated with units of speech that are ca. 2-2.5 seconds long (see Horne et al. [5]).

In addition, to prosodic correlates, it has also been observed that there is often a constituent boundary after 2-2.5 seconds of speech. Thus there would seem to be compelling indications to believe that there exists some kind of timing restriction on speech coding. If this can be proven to be the case via more rigorous testing, it is a restriction that can be very useful in developing algorithms for the parsing of spontaneous speech. In the investigation and analysis of timing restrictions on production units, we are thus making the following basic assumptions:

- A 2-2.5 sec speech production unit can contain internal pauses
- A 2-2.5 sec speech production unit does not contain internal inspirations, i.e. inspirations occur only at the edges of production units
- A 2-2.5 sec speech production unit optimally corresponds to a clause or a constituent.

2. Pauses, inhalations and the internal structure of production units

Breath pauses (i.e. inhalations) are assumed not to occur internal to the 2-2.5 sec. production units. Inhalations are rather assumed to occur only at the edges of speech planning units. Since inhalations occur only at the edges of production units, they can be thought of as anchor points for the division of speech into production units.

In the context of automatic speech processing, the recognition of inhalations can be thought of as the first crucial stage in the chunking of speech into information units. This idea incorporates findings of e.g. Winkworth et al. [11] and Hird and Kirsner [4], who show that inhalations occur predominantly at grammatical boundaries. Thus the relationship between breathing breaks and prosodic phrasing is to be expected (see Tseng [9] for a study on the prosodic labelling of speech using breath-group theory).

Figure 1 shows an example of spontaneous speech illustrating how inhalations can function as anchor points for the segmentation of speech into units corresponding to clauses and smaller constitutents. Assuming that inhalations can be detected and labelled at some initial phase of the speech recognition process, segmentation of speech could then proceed to the left and right of the INHALE-labels in 2-2.5 second intervals, searching for prosodic cues which are known to signal boundaries between clauses. For example, the segmentation algorithm could move 2 seconds to the right of the first inhalation in the speech signal in Figure 1; around that point, one would expect to find a prosodic boundary of some kind. The L% tone followed by a silent pause is such a cue which could in its turn be expected to correlate with a syntactic constituent boundary. At this point,

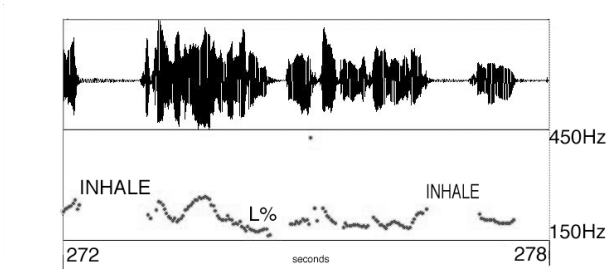

Figure 1: *An example of spontaneous speech illustrating how inhalations (labelled INHALE) can be used as anchors in the segmentation of speech into processing units. The speech between the inhalations consists of two clauses:* Så tränar man med jämna mellanrum *'So you train regularly' and* det gick pågick ju under flera månader *'it last lasted for several months'. The clauses constitute two prosodic phrases separated by a pause.*

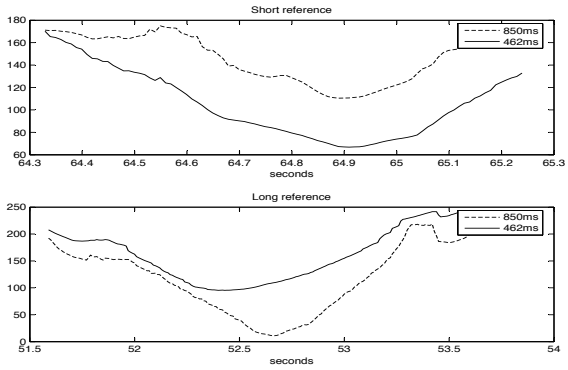

Figure 2: *Distances from test sounds when searched for themselves.*

then, one could insert a production unit boundary after the L% tone.

In summary, inhaling is assumed to play an important role in the delimitation of speech production units: Inhalations only occur at edges and can thus function as anchors for the grouping of speech into 2-2.5 sec speech chunks. Local prosodic information (pauses, boundary tones (H%/L%) and the timing restriction, can be used to make a further segmentation of spontaneous speech into 2-2.5 sec production units.

Within the area of speech technology, relatively little attention has been given to the use of information on breathing. Sundaram and Narayanan [8] have, however, shown that including breathing pauses in their model of speech synthesis increased the naturalness of the synthetic speech. Within the area of speech recognition, Weilhammer and Schiel [10], using hand-labelled data, have shown that using information on breathing improves test set perplexity and recognition accuracy. Breathing is assumed to contribute linguistic information related to its position in speech. Their results suggested a need to improve the acoustical modelling of breathing. The current study has had as its goal to develop a method that would allow one to automatically detect inhalations.

3. Method

Inhalations are characterized by noise and the lack of fundamental frequency, features that they share with voiceless fricatives. They can also have a formant structure if the inhalation is oral. Their duration varies considerably; sometimes it is difficult to identify inhalations both visually in the waveform and auditorily.

Our method of finding inhalations in speech is similar to the template matching technique sometimes used in isolated word recognition. In our approach, a distance measure is calculated between a reference sound and an equally long portion of the sound under examination. A low distance value means that the two compared sounds are likely to be similar. By using an inhalation as the reference sound, this method effectively locates other sound portions that are similar to this inhalation.

Since inhalations were observed to exhibit formant structure, we decided to use a feature representation consisting of the standard Mel Frequency Cepstral Cofficients (MFCC), obtained by performing a Discrete Cosine Transform of the mel-scaled filterbank spectrum. MFCCs were calculated for every 5 ms. The comparison is then done by computing the euclidean distance between the cepstral coefficents of each frame of the two sounds. This results in a matrix with local distances from which the global distance may be calculated using Dynamic Time Warping (DTW). The DTW effectively finds the minimum global distance between the two sounds by using the local distances. The only requirement is that the endpoints of the two sounds coincide. Different temporal structure between the reference and the examined sound is, however, allowed.

The portion of the examined sound is then shifted 10 ms to the right, and the process is repeated. We thus get a distance vector, where each value in the vector represents the total cost between the reference sound and a portion of the examined sound. In this study we used two reference inhalation sounds, taken from a female speaker. The two sounds had different durations, 850 and 462 ms, respectively. The MFCC and DTW analysis were performed using Praat [2] and the distance vector was saved to a textfile for further processing.

The resulting textfile containing the distances between the examined sound and the reference sounds is imported into `Matlab` for further analysis. The Analysis extracts the parts of the distance vector where the distance is lower than a threshold, indicating that an inhalation is present.

The need for a threshold value is quite obvious when applying the algorithm in a search for the reference sounds themselves. In figure 2 one can clearly see that even though the examined sound is the reference sound itself, the algorithm does not produce a value of zero (indicating a perfect match) at any point. This is due to the fact that the alignment of the reference sound can differ from the examined sound by as much as 9ms.

In figure 3, the distance vectors for both a long (850ms) and a short (462ms) sound are shown. Notice the minima in the two vectors. The plot of the corresponding spectrum of the examined sound in the lower part of the same figure exhibits two distinct inhalations aligned with the aforementioned minimas. The minimas have their lowest points at the *end* of the inhalations, which is to be expected given the implementation of the detection algorithm.

When applying the method on recordings of different speakers and environments, it is apparent that the single most important variable is the *threshold* value.This value is what ultimately governs which parts are to be labeled as inhalations and which are not. A low *threshold* value yields few but accurately identified inhalations and a high value results in many identi-

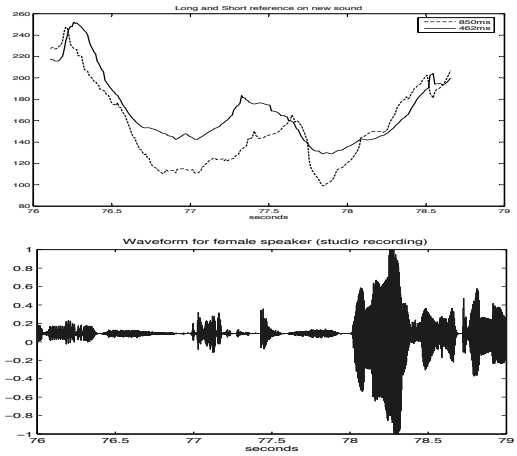

Figure 3: *Distances from test sounds applied on female speaker.*

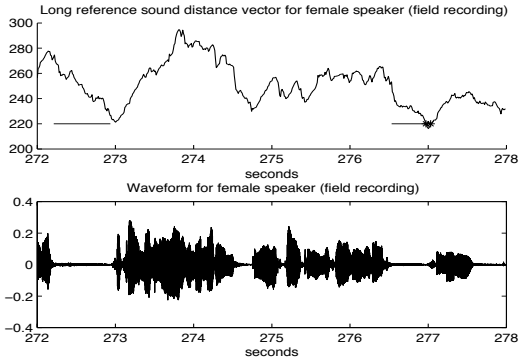

Figure 4: *Distance vector of field recording of female speaker as generated from long reference sound. Stars indicate detected inhalations, lines indicate tagged inhalations. Recording from the SweDia material.*

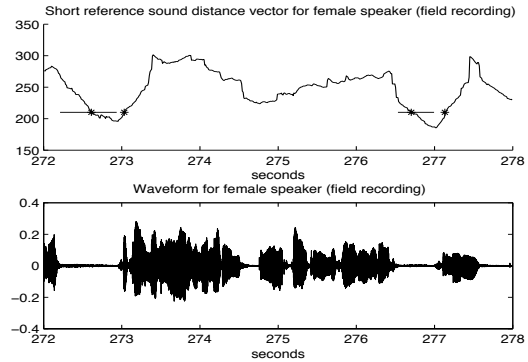

Figure 5: *Distance vector of field recording of female speaker as generated from short reference sound. Stars indicate detected inhalations, lines indicate tagged inhalations. Recording from the SweDia material.*

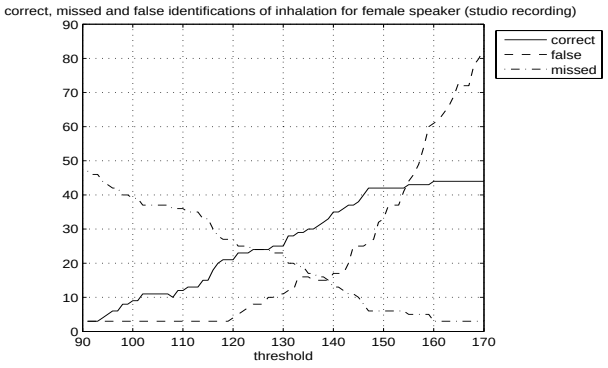

Figure 6: *Number of correct, false and missed identifications of inhalations in studio recording of female speaker as generated from short reference sound.*

fications but at the cost of lower precision. The length of the reference sound is also an important factor as regards detecting inhalations; a longer reference sound performs poorer than a shorter one as can be seen in figures 4 and 5.

4. Results

We used the algorithm to search for speech pauses in three different sets of recordings where the inhalations had previously been tagged manually. In no instance did the algorithm find all of the tagged inhalations. As can be seen in figures 6 to 8, the number of correctly identified inhalations can be relatively high but this also comes at the cost of a large number of false identifications. These false identifications seem to be associated mainly with sounds of the following types: exhalations, word-final aspirated sounds, whispers and voiceless fricatives. To minimize the erroneously identified inhalations, some sort of heuristic must be applied so as not to identify these parts of the speech signal as inhalations.

As the sounds used as references when searching for breath pauses were taken from a female speaker, one might have expected that the algorithm would be most successful when applied to other female speakers. However, this is not the case. Comparing the number of correct, false and missed identifications of inhalations for a male speaker with a female one clearly indicates that the algorithm works better for the male speaker than for the female one (see figures 6 to 8).

As can be seen, the distance vector generated from the short reference sound (figure 5) has three distinct minimas which correspond to the two inhalations present in the speech signal. The longer reference sound (figure 4) exhibits a much more ambiguous set of minimas. Applying the algorithm to noisier recordings illustrates another major problem; the number of false hits increases almost exponentially as the *threshold* value increases (see figure 9). This is probably due to the noisier nature of the signal and the fact that the relative distance between the highest and lowest value in the distance vector is greater than the same distance in the vector generated from the studio recordings.

5. Conclusions and further work

The proposed method to identify breath pauses in spontaneous speech performs fairly well on recordings in a environment where the quality of the recording can be controlled. For noisy signals or signals of low average intensity, the method still can identify breath pauses but at the cost of high rate of erronously identified inhalations. This problem can presumably be addressed by transforming the signal to a normalised space and

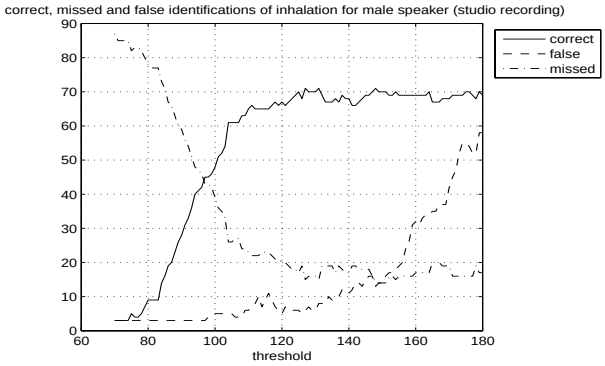

Figure 7: *Number of correct, false and missed identifications of inhalations in studio recording of male speaker as generated from long reference sound.*

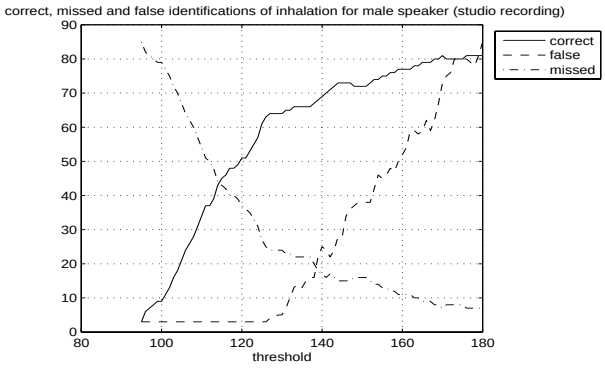

Figure 8: *Number of correct, false and missed identifications of inhalations in studio recording of male speaker as generated from short reference sound.*

thus minimise the effect of noise and low intensity in the signal and is something that we would like to investigate further.

The idea on which the method is based, i.e. that the inhalation signal exhibits properties that are always present, seems to hold and can be utilised in an automatic inhalation detector. These properties do not seem to be dependent on gender, a fact that is most promising, since we then only need to consider one set of properties when constructing a detector for inhalations.

Towards the end of the investigation it was suggested (Franz Clermont, personal communication) that one could use solely the information from the cepstral coefficients in order to develop a direct method for detecting inhalation pauses. The idea is to use sums of the squares of subsets of cepstral coefficients, thus utilizing the fact that the higher order coefficients indicate the presence of F0, while the lowest are related to the energy of the signal. Preliminary results from this approach appear promising and are currently being investigated in more detail.

6. Acknowledgement

This study has been supported by grant 2001-06309 from the VINNOVA Language Technology Program.

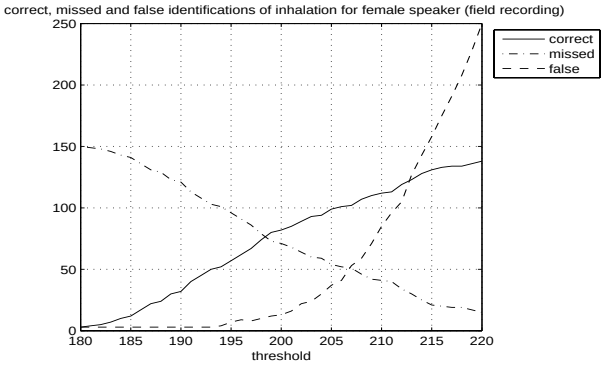

Figure 9: *Number of correct, false and missed identifications of inhalations in field recording of female speaker as generated from short reference sound.*

7. References

[1] Baddeley, A. Human Memory: Theory and Practice. Hove: Psychology Press, 1997.

[2] Boersma, P. and Weenink, D. Praat: doing phonetics by computer (Version 4.3.02) [Computer program]. Retrieved February 27, 2005 from http://www.praat.org/, 2005.

[3] Fant, G. and Kruckenberg, A. “On the quantal nature of speech timing”, ICSLP 96, 2044-2047, 1996.

[4] Hird, K. and Kirsner, K. “The relationship between prosody and breathing in spontaneous discourse”, Brain and Language 80, 536-555, 2002.

[5] Horne, M., Frid, J., and Roll, M. “Timing restrictions on prosodic phrasing”, Nordic Prosody IX. Frankfurt am Main: P. Lang, 2005.

[6] Levelt, W. Speaking: From Intention to Articulation. Cambridge, Mass.: MIT Press, 1989.

[7] Sigurd, B. “How to make a text production system speak”, Working Papers (Dept. of linguistics, U. of Lund) 25, 179-194, 1983.

[8] Sundaram, S. and Narayanan, S. ”An empirical text transformation method for spontaneous speech synthesizers”. Eurospeech 2003.

[9] Tseng, C. ”The prosodic status of breaks in running speech: examination and evaluation”, Speech Prosody 2004.

[10] Weilhammer, K. and Schiel, F, ”Investigations of language structure by means of language models incorporating breathing and articulatory noise”, International Conference of Phonetic Sciences, 1999.

[11] Winkworth, A., Davis, P., Adams, R. and Ellis, E. “Breathing patterns during spontaneous speech”, Journal of Speech and Hearing Research 38, 124-144, 1995.