Speech identification using a sequence-based heuristic

Gregor Heinrich
Fraunhofer Institut f. Graphische Datenverarbeitung, Fraunhoferstr. 5, 64283 Darmstadt, Germany
E-mail: *gregor.heinrich@igd.fhg.de*

Abstract - *This article shows an approach to identify speech in a given signal with moderate noise levels. It is based on the segmental structure of speech processes and has two advantages: It is trained exclusively from speech sounds and has a relatively low computational footprint at testing time.*

Keywords - *Speech detection, signal classification, perceptual linear prediction, support vector machines.*

1. INTRODUCTION

Detecting speech in their surroundings is considered an easy task by human listeners: We are still able to perceive the presence of speech even if its information is lost, for instance because of strong simultaneous noise or when an unknown language is spoken. In technical applications, this detection of a signal's "speechiness" is useful for instance in multimedia databases, where the type of a signal is automatically determined for metadata generation and content-based indexing, or in front-ends to automatic speech recognition systems, where the acoustic environment is scanned for speakers to recognize. Similar requirements exist in hearing aids and ambient intelligence applications.

Early work in speech/non-speech signal classification was accomplished, e.g., by Hoyt and Wechsler [1]-[3], and, more specifically for speech/music discrimination, by Scheirer and Slaney [4]. This research and most later work used a general architecture of signal classifiers: The incoming signal is segmented into frames, discriminative features are extracted and finally classified.

Many features for speech detection and the more general problem of sound classification stem from the speech processing area, such as the representations of spectral envelope, mel-frequency cepstral coefficients (MFCC) [5] and perceptual linear prediction (PLP) coefficients [6], and time-domain features, such as short-time energy and zero-crossing rate. With the more recent developments in multimedia databases and multimedia content description standards like MPEG-7, new standardised feature sets have become increasingly important. In MPEG-7, these so-called descriptors are dedicated to capture the properties of more general sounds, but with some focus on speech and music. Signal classification on the basis of MPEG-7 feature sets, which include, e.g., spectral envelope coefficients, spectral centroid, flatness and spread has been done, e.g., by Kim et al. [7] and Xiong et al. [8].

More recent work on signal classification is that of Lu et al. [9], where a very good recognition rate is achieved using MFCC and a spectral flux feature [4], among others. Classification with support vector machines reportedly achieved a recognition rate beyond 99%.

Most work was achieved using static signal properties and their derivatives. Some alternative approaches use the explicit sequence structure of the signals to discriminate them. Recent work has been done by Ajmera et al. [10], who segment audio streams into speech and music using specific hidden Markov models conditioned on the segment length. An earlier approach introduced by Hoyt and Wechsler [1] uses the curviness of formant trajectories over a longer period to detect speech. Scheirer and Slaney's 4Hz modulation energy feature [4] uses the rate of syllables; vowels in continuous speech exert a peak in short-time energy at an approximate rate of 4Hz.

The concurrently developed sequence model by this paper's author [11] exploits the same syllabic properties but extends this by a classification of the signal segments.

This paper introduces a sequence heuristic that augments the one described in [11] by an improved classification algorithm. The idea is introduced in Section 2, and Section 3 proposes an algorithm implementation whose recognition results are outlined in Section 4.

2. A SEQUENCE-BASED HEURISTIC

Speech is a process continuously varying in time, and one way to account for this non-stationarity is to consider speech a temporal sequence of an alphabet of states. The idea of the proposed speech detection algorithm is that non-stationary non-speech signals are very likely to generate sequences different from these found in speech when using the same alphabet, while stationary non-speech signals are trivial to identify.

The segments of speech considered here are phonemes, which in continuous speech are hierarchically combined to syllables, words, phrases and so on. The most important property for

sequence-based detection is the syllabic structure: Continuous speech alternates between vowels as the central parts of syllables with high frame energy and phonemes with different levels of lower frame energy and different spectral properties. If a detection model is based on phonemes, consequently vowels are most useful to be incorporated into a sequence model: A classifier could just look for vowel-like segments in certain temporal distances. Scheirer and Slaney [4] in principle adopted this in their 4Hz modulation energy measure.

An approach beyond this vowel-modeling is to partition the feature space into a more complete set of phoneme classes and classify lower-energy segments, as well. Such a partition motivated by spectral similarities is the basis for the classification algorithm in the next section.

3. CLASSIFICATION ALGORITHM

The experimental realisation of the speech sequence heuristic comprises the three steps: (1) frame-level feature extraction, (2) sequence generation using segmentation and classification, and (3) the heuristic sequence model test (Fig. 1).

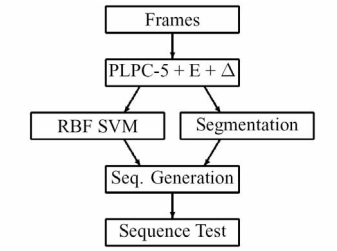

Fig. 1. The architecture of the speech detection approach.

In the initial feature extraction step, the signal is pre-emphasised and windowed into frames of 23ms duration with a half-frame overlap [5]. As classification features, perceptual linear prediction cepstral (PLPC) coefficients of order 5 have been chosen because of good speaker independence [6], compact and de-correlated representation of the signal and computational efficiency.

Next, the frame-based signal must be transformed into a segment-based sequence with a class-label assigned to each segment. This is done by combining points of maximal acoustic change (acoustic landmarks) with the boundaries of classes, using segmentation and classification information.

As a basis for the detection of acoustic landmarks, a Euclidean distance measure between the cepstral coefficients, $\vec{c}_i$, of frames with a certain time difference, $d_i = \| \vec{c}_i - \vec{c}_{i-k} \|_2$, is employed. In the present case, the time difference is set to approx. 11ms (i.e., $k = 1$). The peaks of d_i can be interpreted as points where the signal spectrum exhibits more significant changes compared to the vicinity, such as phoneme boundaries or instants of rapid spectral change in interferences. Measures like d_i commonly produce a great number of peaks which are not easily interpretable. These insertion errors can be reduced by smoothing: The frame-to-frame Euclidean distance is convolved with a "Mexican hat" function, the second derivative of a Gaussian, with standard deviation set to 10 ms. The peaks of the resulting acoustic change function are interpreted as acoustic landmarks and used for segmenting the waveform if they exceed a threshold (see Fig. 2).

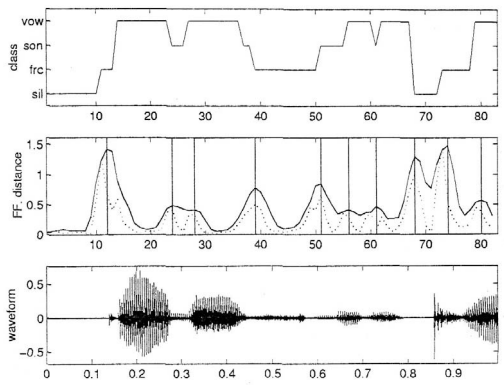

Fig. 2. Segmentation/classification of the beginning of the utterance “Don’t ask me to carry ... ” by a female talker. The middle plot shows the original frame-to-frame Euclidean distance (dotted) as well as the change function where the acoustic landmarks are derived from.

Classification of each frame is performed with a support vector machine (SVM) [12] using the one-against-all multi-classification approach and an RBF kernel. SVMs are able to discriminate classes via topologically complex hyperplanes and are computationally efficient at testing time. The target classes are {vowel, sonorant-consonant, fricative, silence}, i.e., classes of spectrally similar phonemes and a null class. Because the transient character of plosives cannot be handled reliably, separate rules for this class are introduced.

To generate a sequence of segment data, the information of the acoustic landmarks and class labels is combined using the following method:

- Find class regions, which are intervals in the signal during which the class-label does not change and which are longer than 50ms.
- Find acoustic landmarks within the class region and divide the class region into segments at each detected boundary.
- Find class regions which are shorter than 25ms. *For each* of these transient regions:
- *If* (1) the change function across the boundaries of the region is lower than a threshold *and if* (2) their surrounding regions belong to the same class: Merge short regions with surrounding regions. *Otherwise* collect short regions in a

class of transient segments that do not contribute to the estimation of sequence statistics.

This procedure generates a value triple for each segment containing phoneme class, duration and average energy.

Before the signal is passed on to the sequence model classifier, an initial test is performed that classifies observation intervals as non-speech which (1) do not include vowel segments or (2) include only segments of one class and/or silence. This effectively rejects stationary signals and non-stationary signals with non-vowel spectra.

3.1. Sequence Classification

The sequence model classifier as the final part of the algorithm is designed to collect several scores for the segments within an observation interval that correspond to discriminative properties of speech, namely segment durations, vowel/consonant energy ratio and vowel-vowel center separation time.

The first score is determined from the duration d_n of each segment and its class ω_n, $s_{\text{dur}}(n) = S(d_n \mid \omega_n)$, which is a lookup in a class-dependent duration histogram. The duration score is accumulated by a weighted average of the segment-duration scores within the observation interval. The weighting is proportional to the segment durations and additionally emphasises the importance of vowel segments by a constant α:

$$s_1(m) = \frac{\sum_{i<m} d_i^{\text{cons}} S(d_i^{\text{cons}} \mid \omega_i) + \alpha \sum_{j<m} d_j^{\text{vow}} S(d_j^{\text{vow}} \mid \omega_j = \text{vow})}{\sum_{k<m} d_k^{\text{cons,vow}}} \quad (1)$$

The result s_1 can be imagined as a frame-based average measure over the observation interval ($i,j,k<m$) which is preferable to segment-based averages in order to avoid a bias towards short segments. The two other scores are computed for the entire duration of the observation interval:

- The scores of the average duration between the center points of adjacent segments, which are classified as vowels and exceed a duration of 30ms, $s_2 = S(\overline{d_{\text{vv}}})$. *If* only one vowel segment is contained in the sample, the vowel-vowel separation does not contribute to the score, and the false alarm probability is increased. In order to use the same decision threshold as in the normal case, the score is set to the average vowel-vowel separation score obtained for speech, i.e., $s_2 = \text{E}\{s_2\}$.
- The probability of the vowel/consonant energy ratio averaged over the observation interval, $s_3 = P(r_{\text{vow,cons}})$, which is modeled Gaussian.

In the next processing step, the three partial scores for the observation interval are combined to a total score that is defined by a gamma operator, $s = \left(\prod_{i=1}^{3} s_i\right)^{1-\gamma} \left(1 - \prod_{i=1}^{3} 1 - s_i\right)^{\gamma}$.

The weighting factor γ controls the balance between AND and OR behavior of the gamma operator and can be used to adjust the false-alarm robustness. With the gamma operator, a constraint $s_i \in [0,1]$ must be ensured.

The actual decision is made by thresholding the total score, $s(m)$. The index *m* refers to the number of the last full segment recognised in the observation interval. The size of the observation interval must be chosen according to two criteria: (1) the time allowed for a decision, and (2) the confidence of the decision. It can be increased with an increasing number of available segments. Then the confidence of decision is low at the beginning of an utterance and increases when more and more frames/segments are input to the system. In the current version, the observation interval reaches across the entire input file.

3.2. Training

Training is performed in two stages. First, the SVM phoneme-class classifier is trained with labeled speech samples from different talkers. According to the quality of the training labels, the training error robustness of the SVM is set. Subsequently, the distributions for the sequence model can be estimated by unsupervised training, i.e., the training relies on the behavior of the segmentation/classification stage. The statistical models for the phoneme-dependent durations are trained non-parametrically, i.e., histograms for the class durations are collected and smoothed. For the vowel/consonant energy ratio, a simple estimation of the Gaussian mean and variance is performed. As a final step, the weights α and γ are empirically chosen at the end of the training phase.

4. EVALUATION

Training was done on 4500 frames from 40 different talkers from the TIMIT corpus for the classifier, durational statistics and energy ratios. Plosive sounds were left out (they tended to be classified as sonorants and fricatives; cf. Fig. 2, e.g., frames 13–15 and 47–50). The durational distributions measured for the three remaining phoneme classes are shown in Fig. 3 and show shortened vowel durations due to some misclassifications at boundaries between vowels and sonorant consonants. For the duration score and the final class decision, the vowel weighting constant was set to $\alpha = 2$ and the AND/OR weight was set to $\gamma = 0.7$, emphasising OR behavior. The scores have in addition been observed separately, see Fig. 4.

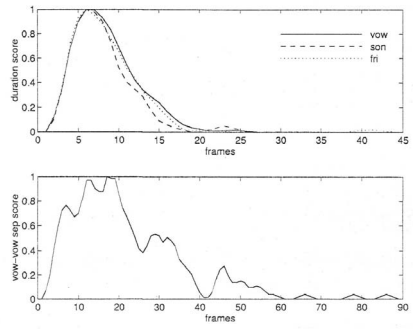

Fig. 3. Normalized duration histograms for the three phoneme classes (top) and vowel-vowel separation (bottom) used by the classifier (1 frame = 11ms).

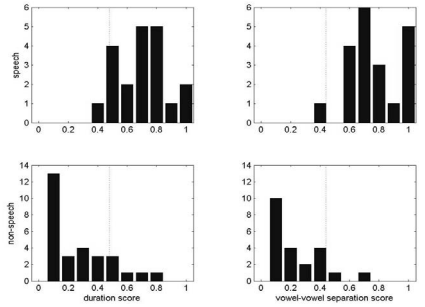

Fig. 4. Histograms of the two scores duration and vowel-vowel separation.

The algorithm has proven the general ability to identify the speech samples with a set of 20 speech samples and 30 structured nonspeech samples including music and crowd noise. With the optimal thresholding and weighting parameters, the recognition rate was >90% with the complete files as observation interval (2–5sec). Speech samples are classified generally reliably with the durational values for read continuous speech. Harmonic non-speech sounds like music were often classified constantly as vowels or sonorants and thus could be rejected. Rejection of short noise sounds like keyboard noise or crackling fire were classified as transients and thus were rejected with the transient / short segments rule that excludes the sounds from classification.

5. CONCLUSION

A speech detection approach has been proposed that classifies speech and non-speech signals using a combination of heuristics based on the syllabic structure of speech. The model was developed primarily under the working assumption of non-simultaneous noise, but by appropriate feature selection (e.g., harmonics in vowels), some robustness to noise can be expected. Compared to HMM-based sequence algorithms, no expensive model decoding procedure is necessary.

Future research should concentrate on the behaviour of the algorithm with limited observation intervals, variable speaking rates and the possibility to segment an audio stream into speech and non-speech regions. Further, explicit handling of plosive phonemes / transient non-speech sounds can improve the model.

ACKNOWLEDGEMENTS

The presented work has been partially funded by Sensimetrics Corp., Somerville, Massachusetts, as well as the European Commission.

REFERENCES

[1] J.D. Hoyt and H. Wechsler, “Detection of human speech in structured noise”. In *Proc. ICASSP-94*, Vol. II, 1994, pp. 237–40.

[2] J.D. Hoyt and H. Wechsler, “Detection of human speech using hybrid recognition models”. In *Proc. ICASSP-94*, Vol. I, 1994, pp. 330–3.

[3] J.D. Hoyt and H. Wechsler, “RBF models for detection of human speech in structured noise”. In *Proc. ICASSP-94*, Vol. V, 1994, pp. 4493–6.

[4] E. Scheirer and M. Slaney, “Construction and evaluation of a robust multifeature speech/music discriminator”. In Proc. *ICASSP-97*, 1997, pp. 1331–1334.

[5] L.R. Rabiner, and B.-H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ: Prentice Hall, 1993.

[6] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech”. *J. Acoust. Soc. Am.* 57(4), Apr. 1990, pp. 1738–52.

[7] H.-G. Kim, J.J. Burred and T. Sikora, “How efficient is MPEG-7 for general sound recognition”. In *Proc. AES 25th International Conference*, 2004.

[8] Z. Xiong, R. Radhakrishnan, A. Divakaran and T. Huang, “Comparing MFCC and MPEG-7 audio features for feature extraction, Maximum Likelihood HMM and Entropic Prior HMM for sports audio classification”. In *Proc. IEEE Int. Conf. on Multimedia and Expo* (*ICME*), 2003.

[9] L. Lu, H.-J. Zhang and S.Z. Li, “Content-based audio classification and segmentation by using support vector machines”. *Multimedia Systems* 8, 2003, pp. 482–492.

[10] J. Ajmera, I. McCowan and H. Bourlard, “Speech/music segmentation using entropy and dynamism features in a HMM classification framework”. *Speech Communication* 40, 2003, pp. 351–363.

[11] G. Heinrich, *Speech detection algorithms.* Diplomarbeit, TU Darmstadt, Jan. 1998.

[12] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press, 2000.