# A NOISE-ROBUST SUBSPACE-BASED SOUND-CLASS DETECTOR

*Wolfgang Wokurek*
*Institut für Maschinelle Sprachverarbeitung,*
*Universität Stuttgart, Germany*

## ABSTRACT

A computationally efficient approach to the automatic segmentation (labeling) of noise disturbed speech is presented. The segmentation algorithm employs short term spectrum based feature vectors and a subspace representation of the sound classes. The two sound classes of vowels and unvoiced fricatives are trained with the TIMIT acoustic phonetic continuous speech corpus. The sound class detector is applied in a speech enhancement system and for the automatic segment duration measurement.

## INTRODUCTION

Originally this automatic sound class detection algorithm was developed as an improved replacement for the speech pause detector of a speech enhancement system [Wokurek 94]. Clearly this application requires noise robustness. Furthermore, a solution with low computational effort was sought to allow real time implementation. Representing the sound classes by subspaces meets both goals.

The speech signal is transformed into a sequence of feature vectors. This transformation controls which properties of the speech signal are represented by the length and by the direction of the feature vectors. In order to distinguish between different speech sounds the transformation will control the direction of the feature vectors by the shape of the spectrum. On the other hand, the feature vectors do not contain any pitch frequency information.

Unfortunately, no transformation is known that converts each phoneme or even each speech sound to a uniquely defined direction within the feature space [Furui 92]. Context, allophonic variations and noise ensure that the feature vector of every speech sound moves around quite a lot in the feature vector space. For automatic speech sound recognition it is necessary to describe the directions of the feature vector that are possible for each speech sound. A standard approach is to collect a number of representative feature vectors for each speech sound. Either the collection of feature vectors or a statistical description of them may be used as a representative of each speech sound. If the collection of feature vectors is used, their number is likely to exceed 1000. That number is met if e.g. every of 50 speech sounds is represented by only 20 vectors.

Is this large number of representative elements inevitably? A vector represents a single direction, in this sense it is a 'small' object in a vector space. A plane is a 'larger' object in that sense



— it contains infinitly many directions. Furthermore it only needs two (orthogonal) vectors to be defined. If even a plane is 'too small', $D$-dimensional subspaces could be used[1].

Is there any problem to represent speech sounds by planes instead of vectors? The problem might be that the plane is likely not only to contain the directions of the speech sound, but many other directions as well. So planes — or $D$-dimensional subspaces — should be used with care, and *not without further evidence.* In the case of this study it is observed experimentally that the feature vectors of the vowels [ieaou] lie in the vicinity of a plane. This motivates the notion of representing sets of speech sounds — sound classes — by subspaces.

Further experiments demonstrate the noise-robustness of a sound-class detector based on that subspace representation. These experiments indicate that the noise-robustness results from broadening the scope of discrimination from sounds to sound-classes. It should be noted however, that the sound-classes may not be defined arbitrarily. Only sounds with similar spectra are efficiently represented by a (low dimensional) subspace.

Finally it is important that the 'online' operation of this subspace-based sound-class detector is computationally efficient. Only the 'offline' training of the subspace representation of the sound-classes is computationally expensive.

## FEATURE VECTORS

The disturbed speech signal is converted to feature vectors employing the short time energies of the output signals of a band-pass filter bank. Let each coordinate of the feature vector represent the short term energy within the bandwidth of a single channel of the filter bank. Then e.g. a formant produces a signal of high amplitude at the output of a certain filter bank channel and pulls the feature vector into it's direction. By this mechanism, the short term spectrum of the speech signal determines the direction of the feature vector. Each change in the formant structure changes the short term spectrum and turns the feature vector.

The short term spectrum of the speech signal is one ingredient to the feature vector direction — the bandwidth design of the filter bank is the second. Only those spectral changes turn the feature vector, that change it's coordinates; i.e. a formant movement turns the feature vector if and only if the formant moves into a different channel of the filter bank. Therefore different bandwidth designs of the filter bank are used.

The basis is the constant bandwidth of each channel. Here 20 channels with a bandwidth of 400 Hz are used to cover the frequency band from 0 to 8 kHz. A linear mapping occurs between the number of each channel and it's center frequency, therefore it is addressed as 'linear filter bank design'.

In contrast to that, a 'bark filter bank design' is employed to represent the psychoacoustic scale of critical frequency bands. Again, 20 channels cover the frequency band from 0 to 8 kHz, but the bandwidth starts with 100 Hz at low center frequency and increases to more than 1 kHz. Both filter bank designs are implemented using a computationally efficient algorithm — the 'Lerner filter bank' [Doblinger 91] [Lerner 64].

[1] A vector is a 1-dimensional subspace and a plane is a 2-dimensional subspace of the $N$-dimensional vector space.

The dimension of the feature vector space is defined by the number of filter bank channels. Hence the dimension is $N = 20$ for both, the linear and the bark filter bank design.

Given the noisy speech samples $x(n)$, the filter bank and the short term energy measurement of each channel results in the sequence of feature vectors $\mathbf{z}(m)$

$$x(n) \rightarrow \mathbf{z}(m) \quad .$$

The short term energies are computed with a time window duration of 20 ms. Therefore the feature vectors are low pass signals that do not require the sampling rate of the speech signal (20 kHz). The feature vectors are calculated at a sampling rate of 100 Hz, that is significantly lower. Hence the speech class detector operates at the lower sampling rate, what helps to limit the computational load.

## SUBSPACE EXTRACTION

For the purpose of subspace extraction the feature vectors of all speech sounds that will train the considered sound class are collected within the observation Matrix

$$\mathbf{A} = (\ldots, \mathbf{z}(m), \ldots) \quad .$$

Now a subspace is required, that represents all these feature vectors in some sense. Here, the least square minimization of all vector components 'outside' (i.e. orthogonal to) the subspace is used as optimization criterion. A solution to that problem is found by the eigendecomposition of the correlation matrix of all feature vectors

$$\mathbf{C} = \mathbf{A}\mathbf{A}^* = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^* \quad ,$$

where $\mathbf{U}$ is the orthogonal matrix of eigenvectors and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues [Golub 89].

Once the eigenvectors and eigenvalues of $\mathbf{C}$ are computed, the $D$ dimensional subspace is spanned by the eigenvectors that correspond to the $D$ largest eigenvalues. A threshold of 1% of the largest eigenvalue is used for automatic subspace dimension determination.

This algorithm represents each sound class and the noise signal by a single subspace. The automatic sound class detection is performed by comparing the actual feature vector to all subspaces. 'Winner' is the class that contains the largest component of the actual feature vector. Finally, a three point median filter improves the decision by removing isolated deviations.

## SOUND CLASSES

Initially the two sound classes of vowels [i:e:a:o:u:] and unvoiced fricatives [fsçx$\int$] are trained with the TIMIT acoustic-phonetic continuous speech corpus. The subspace extraction results in a 1-dimensional noise subspace, a 2-dimensional subspace for the vowels and a 3-dimensional subspace for the fricatives. The resulting sound-class detector correctly detects the vowels, but confusions between the noise and the fricatives occur.

Due to larger spectral differences within the sound class of fricatives, the subspace of that class tends to 'catch' some of the feature vectors of noise segments. To minimize these errors, a further optimization of the sound classes is applied.

Now the sounds are exchanged between the classes until the smallest angle between every two subspaces is maximized. In addition, a smaller number of sound classes is preferred. The full search over all possible sound class partitions results in three sound classes. Due to better separation of the subspaces, the detection error rate decreases.



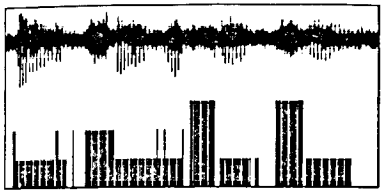

Figure 1: Detection of optimized sound classes: SNR=10dB, white noise

## RESULTS

Figure 1 shows the segmentation of a noisy signal. The German utterance 'Deutscher Übersetzung' is analyzed. It is disturbed with white noise at a signal to noise ratio of 10 dB. The sound class detector employs the three optimized sound classes as well as the noise class.

Class #0 corresponds to the noise signal and is visible by 'missing' marks in Figure 1. Sound class #1 contains the voiced speech sounds. Sound class #2 is evoked by the $\int$, d and t sounds. Finally, the s sound is detected as a member of sound class #3.

The sound class detector is applied in a speech enhancement system to replace the speech pause detector. There, the speech spectrum estimation that is required for the enhancement, is controlled by the speech class decision. During noise segments an additional suppression is applied.

A second application of the automatic sound class detection algorithm is the automatic measurement of segment durations. The segment duration is used to normalize the time axis of fundamental frequency contours.

## CONCLUSION

Classes of speech sounds are represented by low dimensional subspaces. The discrimination between sound classes instead of sounds is one source of the noise robustness of the algorithm; the restriction of sound class definition to sounds of similar spectral shape is the second. Finally, the subspace representation results in computational efficiency.

Even a small number of speech sound examples leads to reasonable results of the sound-class detector. This is due to the fact that a deterministic (nonstatistic) model is used. However, an increased number of different training sound examples improves the speaker independency of the segmentation results.

## REFERENCES

[Doblinger 91] G. Doblinger. An efficient algorithm for uniform and nonuniform digital filter banks. In *IEEE Proceedings of ISCAS'91*, vol. 1, pp. 646 - 649. Singapore, 1991.

[Furui 92] S. Furui, M. M. Sondhi. *Advances in speech signal processing.* Marcel Dekker Inc., New York, 1992.

[Golub 89] G. H. Golub, C. F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, Baltimore, 1989.

[Lerner 64] R. M. Lerner. Band - pass filters with linear phase. In *Proceedings of the IEEE*, vol. 52, pp. 249 - 268. March 1964.

[Wokurek 94] W. Wokurek. *Sprachentstörung unter Verwendung eines Lautklassendetektors.* Dissertation, Technische Universität Wien. 1994.