ON THE ROBUST AUTOMATIC SEGMENTATION OF SPONTANEOUS SPEECH

Bojan Petek[1], *Ove Andersen, and Paul Dalsgaard*

Center for PersonKommunikation, Aalborg University, Denmark

ABSTRACT

The results from applying an improved algorithm in the task of automatic segmentation of spontaneous telephone quality speech are presented, and compared to the results from those resulting from superimposing white noise. Three segmentation algorithms are compared which are all based on variants of the Spectral Variation Function. Experimental results are obtained on the OGI multi-language telephone speech corpus (OGI_TS). We show that the use of the auditory forward and backward masking effects prior to the SVF computation increases the robustness of the algorithm to white noise. When the average signal-to-noise ratio (SNR) is decreased to 10dB the peak ratio (defined as the ratio of the number of peaks measured at the target over the original SNRs) is increased by 16%, 12%, and 11% for the MFC (Mel-Frequency Cepstra), RASTA (RelAtive SpecTrAl processing), and the FBDYN (Forward-Backward auditory masking DYNamic cepstra) SVF segmentation algorithms, respectively.

1. INTRODUCTION

The Spectral Variation Function (SVF) is defined as a correlation measure between successive windows of acoustic observation vectors [3, 10, 11]. Several SVF definitions are referenced in the literature [9], and a number of them have been successfully applied to the tasks of automatic segmentation and speech recognition [3]. However, one of the weak points of SVFs have been found to be its sensitivity to noise (i.e., due to noise, spurious SVF peaks occur in all parts of the acoustic signal).

It has been shown in automatic speech recognition of spontaneous speech that by emphasizing the *relevant* spectral dynamics (e.g., the enhancement of formant transitions) the robustness can be improved. To address this and the noise robustness problems, spectral representation simulating the time-frequency characteristics of the auditory forward masking has been proposed for spontaneous speech recognition [1] where a combination of the forward and backward masking effects [2] has recently yielded excellent performance results on recognition of clean and noisy speech.

The next two sections detail a comparative performance analysis of three segmentation algorithms and their evaluation on the OGI_TS.

[1] Visiting researcher from the University of Ljubljana, Slovenia and postdoctoral scholar of the Slovenian Science Foundation.

2. FBDYN-SVF ALGORITHM

Based on these findings an improved speech segmentation algorithm, called FBDYN-SVF, is proposed to address the noise sensitivity problem of the SVF based segmentation. Here a major extension as compared to the original formulation [3] is that prior to the SVF computation the cepstra are smoothed by forward and backward masking liftering.

2.1. Signal Conditioning

Three SVF-based segmentation algorithms (MFC-SVF, RASTA-SVF, and FBDYN-SVF) were tested in three SNR ratio conditions (original, 20dB, and 10dB). White noise was added to a speech signal to obtain the target average SNR. Speech signals were parametrized using the CUED HTK. The parameters for the mel-frequency cepstra were: Hamming window duration 16ms, frame period 5ms, number of output parameters 16, analysis order 19. The RASTA coefficients were computed from the mel-frequency cepstra using the filter [5]

$$H(z) = 0.1\, z^{4}\, \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}}$$

FBDYN (for a complete description, see [2]) parameters used identical forward $l_k(n)$ and backward $r_k(n)$ masking Gaussian lifters defined by

$$l_k(n) = r_k(n) = \alpha\beta^{n-1} \exp\left(-\frac{k^2}{2(q_0 - \nu(n-1))^2}\right)$$

where n denotes the time delay and k the order of the cepstral coefficient vector component. The values of $\alpha = 0.08$, $\beta = 0.78$, $q_0 = 18$, $\nu = 1$ were used. Since a 5ms frame period was used, the duration of the forward N_f and the backward masking effect N_b was set to $N_f = N_b = 6$, i.e., 30ms. Therefore, the selected values determine a digital filter whose transfer function depends on the order k. Figure 1 shows this relationship for the C_1, C_8, C_{16}. It is observed that the attenuation of slow changes increases as the modulation frequency (defined as the frequency of change of a C_n) decreases. In addition, at lower frequencies the attenuation decreases with an increasing order k. These two properties enable the dynamic cepstrum to attenuate stationary or slowly varying wide band noise.

The SVF function [3] in each of the algorithms used a symmetric

window of 11 frames (i.e., 55ms) whose duration has been set in accordance with the results reported by Furui [4].

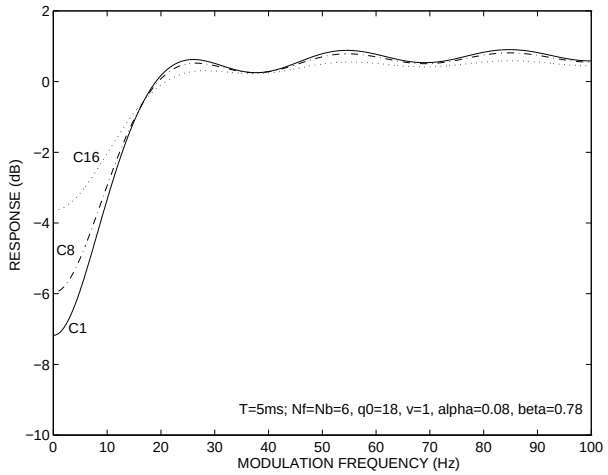

Figure 1: Transfer function of each dynamic cepstral coefficient is order dependent. (see Section 2).

2.2. Performance Evaluation Methodology

Performance evaluation of the speech segmentation algorithms in general is a difficult problem. Problems like inconsistency in methods for manual segmentation complicate the task. Main premises of the adopted evaluation strategy in this work are: 1.) a consistent evaluation of all three segmentation algorithms in regard to the search regions for the SVF peaks, 2.) as sensitive SVF analysis as possible in order to highlight possible difficulties in segmentation of spontaneous, telephone quality speech (i.e., deletion errors), and 3.) the treatment of the oversegmentation as a separate problem using the constraints imposed by phonetic knowledge. Therefore, the segmentation accuracy is defined and evaluated *without* preselected thresholds, i.e., the errors are not classified into the categories gross and fine [6]. In contrast, we describe the deviations in a statistical sense. A similar strategy is also applied in the detection of SVF peaks P_i (Figure 2) where the SVF values are compared over three consecutive frames from which the existence of a peak at frame i is defined, if the values $SVF_{i-1} < SVF_i \geq SVF_{i+1}$.

The oversegmentation ($o.s.$) is defined by [6]

$$o.s. = \left(\frac{\text{TOTAL NUMBER OF SVF PEAKS}}{\text{TOTAL NUMBER OF LABELS}} - 1 \right) 100 \quad [\%]$$

For example by an oversegmentation rate of 210% the number of SVF peaks is 3.1 times the number of (manually positioned) labels measured across the whole database.

Figure 2 illustrates the adopted methodology in assigning the SVF peaks to the manually positioned boundaries. Consider a hypothetical sentence which consists of three manually positioned labels A, B, and C. Two search regions can be established where transitory peaks can be searched for. The boundary for a search region is set to the middle points of the manually positioned labels. If no peaks of

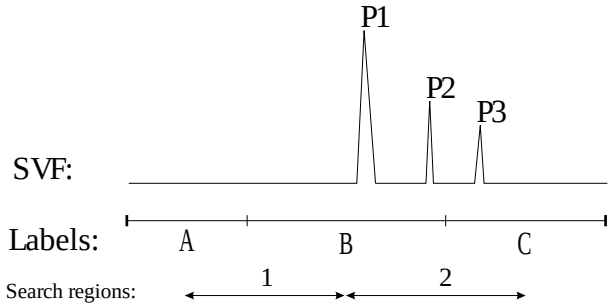

Figure 2: An illustration of the performance evaluation methodology. SVF: Spectral Variation Function, Labels: manually positioned reference segmentation, Search regions: regions where peaks of the SVF are searched for (see text).

the SVF can be found in a search region, a deletion is detected (e.g., in search region 1 of Figure 2). The SVF peak found nearest to the label boundary is marked as a transitory peak (e.g., P2 in Figure 2). If there is a tie, i.e., two peaks occur at the same distance left and right of the transition boundary, the peak to the left of the boundary is marked as transitory. All other peaks in the search region are marked as insertion peaks (e.g., P1 and P3 in Figure 2). Deletions are labeled according to the identity of the search region (e.g., the deletion in the search region 1 is labeled as an **A+B** deletion). Like deletions transitory peaks are labeled similarly, therefore, the transitory peak P2 is labeled as a **B+C** transitory peak. Insertions, however, are labeled according to the label identity (e.g., peaks P1 and P3 are labeled as a **B** and **C** insertions, respectively).

3. EXPERIMENTAL RESULTS

The OGI multi-language telephone speech corpus (OGI_TS) was chosen for the evaluation of the segmentation algorithms. Only the English part of the database, i.e., a subset of 146 files with 81,511 manually positioned Worldbet labels and 117 minutes of speech was used in the experiments reported here.

3.1. Characteristics of the OGI_TS

The selected part of the OGI database was first analyzed for the original SNR. In order to enable an analysis on broad phonemic classes of consonants (**C**), diphthongs (**D**), vowels (**V**) and non-speech (**N**) [7], the 215 unique Worldbet segments were first redefined by removing the diacritics and merging (.pv labels with .unk, and ng_= with N) into a total of 80,302 labels (60 unique). Figure 3 summarizes the relative frequencies of transitions between the broad phonemic classes as given within the database.

3.2. Broad Phonemic Analysis Results

Each of the segmentation algorithms (MFC-SVF, RASTA-SVF, FBDYN-SVF) was evaluated using the methodology described in Section 2.2 at three different SNR conditions (original, 20dB, 10dB). The overall relationships between deletion, insertion, transition statistics of peaks for each of the algorithms across the three SNRs were found to be *qualitatively* similar. Therefore, only the results obtained at the original SNR are going to be presented here.

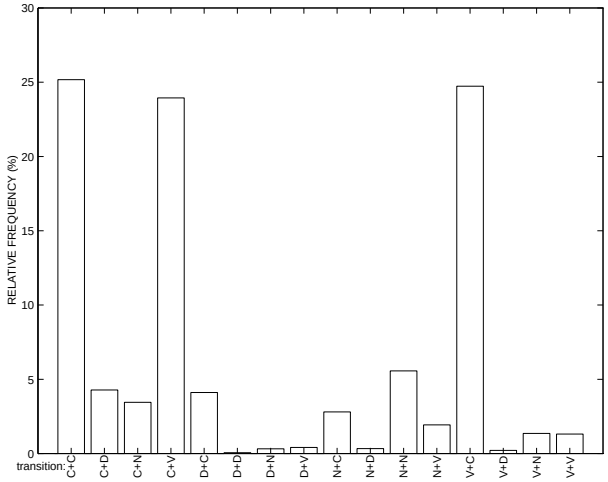

Figure 3: Relative frequencies of transitions between the broad phonemic classes in the selected English OGI_TS.

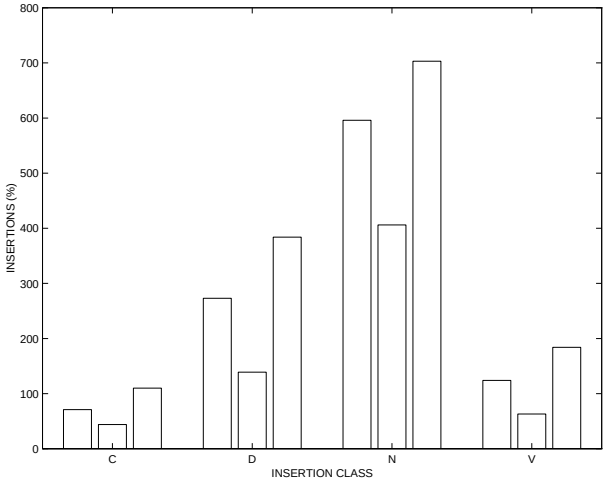

Figure 5: Percentage insertions between the broad phonemic classes at the original SNR. The sequence of algorithms shown is MFC-SVF, RASTA-SVF, FBDYN-SVF for each class, respectively.

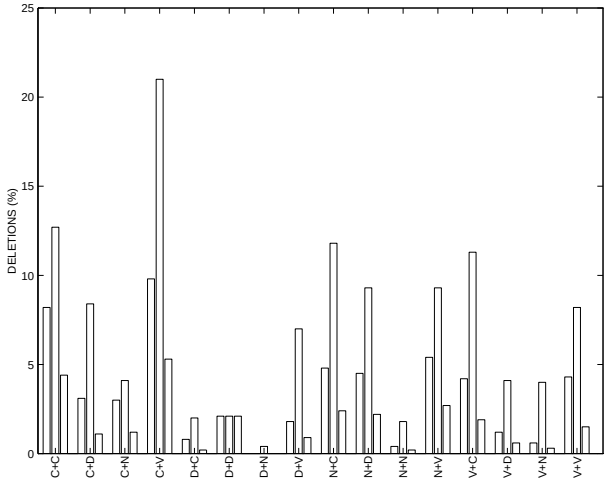

Figure 4: Percentage deletions on transitions between the broad phonemic classes at the original SNR. The sequence of algorithms shown is MFC-SVF, RASTA-SVF, FBDYN-SVF for each class, respectively.

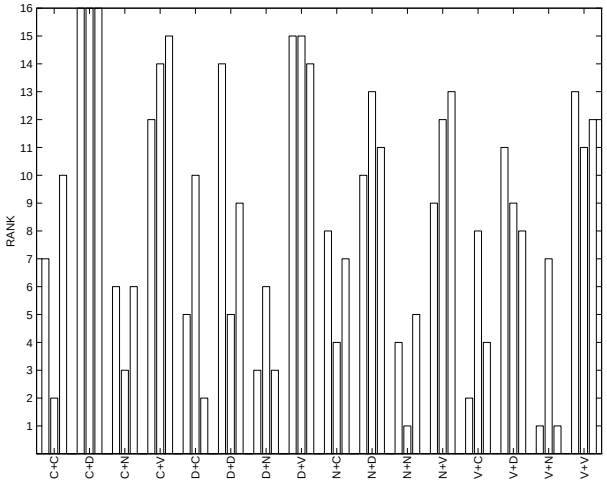

Figure 6: Rank order statistics of the segmentation accuracy as compared to the manually positioned boundaries. Ranks are associated with the average distances between the manually positioned labels and the transitory SVF peaks *within* individual segmentation algorithms. A lower value means better precision. The sequence of the algorithms shown for each class is MFC-SVF, RASTA-SVF, FBDYN-SVF.

Figure 4 shows that the FBDYN-SVF had significantly less deletions than the other two segmentation algorithms, but at the expense of more insertions (Figure 5). For all the algorithms it is observed that the most problematic are the **C+V, C+C, N+C**, and **V+C** transitions.

The **N**-class was found to be the most difficult in view of the number of insertions (Figure 5). In general, the overall very high insertion rates could be explained by the fact that no explicit thresholding was used in the SVF peak detection strategy (see Section 2.2).

Figure 6 presents a comparative (within-algorithm and within-class) rank order statistics of the precision of segmentation. It is observed that transitions **C+D, D+V**, and **C+V** are the least precisely detected for all the algorithms.

3.3. Robustness to Noise

Comparative performance analysis results for the different SNRs are summarized in Table 1. It is observed that for the FBDYN-SVF, peak boundaries on the average are positioned less than two frames apart from the manually positioned boundaries for all SNRs, but at the expense of a high oversegmentation. On the other hand, the RASTA-SVF algorithm yielded the lowest oversegmentation at the expense of a much lower accuracy and deletion problems (e.g., a more than 20% deletion rate for **C+V** class in Figure 4). In general it was found that algorithm accuracies were not heavily affected by the

addition of white noise. Overall the results show that the FBDYN-SVF was found to be more accurate than the RASTA-SVF but at the expense of higher oversegmentation.

	MFC-SVF			
SNR	mean	std	peaks$_t$	o.s.
original	1.73	1.47	75,302	168%
20dB	1.69	1.45	75,282	180%
10dB	1.64	1.39	76,293	210%
	RASTA-SVF			
SNR	mean	std	peaks$_t$	o.s.
original	3.14	1.89	70,197	96%
20dB	3.12	1.87	70,589	103%
10dB	2.92	1.83	72,333	121%
	FBDYN-SVF			
SNR	mean	std	peaks$_t$	o.s.
original	1.55	1.21	77,683	232%
20dB	1.50	1.20	77,643	241%
10dB	1.44	1.15	78,136	270%

Table 1: Accuracy of the SVF-based algorithms for the original and two SNRs. Shown are the average distances and standard deviations (measured in units of 5 ms frames) of the nearest SVF peaks relative to the manually positioned boundaries. Column peaks$_t$ shows the total number of peaks marked as transitory on which the performance scores were calculated.

Robustness of the algorithms to noise is summarized in Table 2. The results show that, e.g., at 10dB SNR the FBDYN-SVF and RASTA-SVF had comparable sensitivity to noise (i.e., 11% and 12% increases in the number of peaks, respectively). On the other hand, the MFC-SVF was observed to be the most sensitive to noise as the number of peaks increased by 16%.

	MFC-SVF	
SNR	peak ratio	deletion ratio
20dB	1.04	1.0
10dB	1.16	0.8
	RASTA-SVF	
SNR	peak ratio	deletion ratio
20dB	1.03	0.96
10dB	1.12	0.79
	FBDYN-SVF	
SNR	peak ratio	deletion ratio
20dB	1.02	1.02
10dB	1.11	0.82

Table 2: Robustness of the segmentation algorithms to noise. Shown are the ratios between the number of peaks (or deletions) detected in noisy over the original (telephone quality) speech signals.

4. CONCLUSIONS

Comparative performance analysis of the MFC-SVF, RASTA-SVF, and FBDYN-SVF segmentation algorithms showed a similar qualitative behaviour on segmentation of the broad phonemic classes. The total number of SVF peaks (the oversegmentation ratio), deletions and insertions are clearly highly correlated for each of the algorithms. Thus in general, more peaks (higher oversegmentation) means less deletion errors and more insertions. On the other hand, the robustness to noise of the FBDYN-SVF and RASTA-SVF algorithms were found to be comparable and superior to the originally formulated SVF-based segmentation algorithm, MFC-SVF.

A systematical study of the sensitivity of the SVF segmentation to other signal processing parameters (e.g., Hamming window length, the HTK analysis order p, the number of cepstral coefficients in the SVF computation) has already been completed as well. Ratios of the SVF mean values for transitory peaks in noisy versus the original SNRs for each of the algorithms have also been analyzed. Due to lack of space, these results together with their implementation in automatic speech recognition are going to be reported in the future.

5. ACKNOWLEDGEMENTS

Bojan Petek gratefully acknowledges a postdoctoral scholarship awarded by the Slovenian Science Foundation. His research was also supported in part by the postdoctoral project Z2-7171-0781-95 granted by the Ministry of Science of Slovenia.

6. REFERENCES

1. Aikawa K., Singer H., Kawahara H., and Tohkura Y., "A Dynamic Cepstrum Incorporating Time-Frequency Masking and its Application to Continuous Speech Recognition", Proc. IEEE ICASSP'93, II:668-671, 1993.
2. Beppu T., and Aikawa K., "Spontaneous Speech Recognition Using Dynamic Cepstra Incorporating Forward and Backward Masking Effect", Proc. ESCA Eurospeech'95, 511-514, 1995.
3. Flammia G., Dalsgaard P., Andersen O., and Lindberg B., "Segment Based Variable Frame Rate Speech Analysis and Recognition Using a Spectral Variation Function", Proc. ICSLP'92, 983-986, 1992.
4. Furui S., "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", IEEE Trans. on ASSP, **34**(1), 52-59, February 1986.
5. Hermansky H., Morgan N., Bayya A., and Kohn P., "RASTA-PLP Speech Analysis Techniques", Proc. IEEE ICASSP'92, 121-124, 1992.
6. Kvale K., *Segmentation and Labelling of Speech*, PhD Dissertation, The Norwegian Institute of Technology, 1993.
7. Lander T., and Metzler S. T., *The CSLU Labeling Guide*, Oregon Graduate Institute report CSLU 003, 1994.
8. Muthusamy Y. K., Cole R. A., and Oshika B. T., "The OGI Multi-Language Telephone Speech Corpus", Proc. ICSLP'92, 895-898, 1992.
9. Nouza J., "A Study on Spectral Variation Functions Applied to Speech Signals", Final report on Project No. 4678, CPK, Aalborg University, June 1994.
10. Svendsen T., and Soong F. K., "On the Automatic Segmentation of Speech Signals", Proc. IEEE ICASSP'87, 77-80, 1987.
11. Wilpon J. G., Juang B. H., and Rabiner L. R., "An Investigation on the Use of Acoustic Sub-Word Units for Automatic Speech Recognition", Proc. IEEE ICASSP'87, 821-824, 1987.