Phone-based Plosive Detection

Andreas Madsack, Grzegorz Dogil, Stefan Uhlich, Yugu Zeng and Bin Yang

Abstract

We compare two segmentation approaches to plosive detection: One aproach is using a uniform segmentation of the speech signal to 10 ms slices whereas the other assumes additional information about the start and end of each phone and uses these values as segmentation boundaries. We show that including this information yields significantly better results than using a uniform segmentation. We test both approaches in three different experiments using the TIMIT corpus: plosive vs. non-plosive recognition, voiced vs. unvoiced plosive detection and individual plosive classification.

Index Terms

Plosive detection, Segmentation, Pattern recognition

I. Introduction

The purpose of this technical report is to present a statistical classification framework for plosive detection. In contrast to the traditionally applied methods which use information of the signal before and after the relevant time frame [1] we perform a decision for each individual speech segment. Our approach can be summarized as follows: First, segment the speech signal into blocks of either a fixed size of 10 ms or a variable size that exploits additional information on the start and end of each phone, second perform a decision for each segment to which class it belongs.

A segmentwise approach does not require training a HMM, which is computational demanding. We can use simple classification schemes such as the Bayes classifiers [2] which are more efficient. However, there is a clear disadvantage to our phone-driven approach. Plosives are phonetically not uniform segments.

A. Madsack and G. Dogil are with the Institute of Natural Language Processing, Universität Stuttgart, Germany
S. Uhlich and B. Yang are with the Chair of System Theory and Signal Processing at the Universität Stuttgart

This work is supported in part by the Deutsche Forschungsgemeinschaft (Collaborative Research Center SFB 732: Incremental specification in context)

To the contrary, they consist of two seperate temporal phases: a silence phase at the beginning (e.g. the average length in TIMIT is 57.1 ms) which is followed by the burst and the release phase (e.g. the average length in TIMIT is 38.5 ms). HMM-based approaches are by nature more suited to cope with this temporal characteristic. Our goal is therefore to find a simple but effective classification method which works on a phonetic basis and yields at least the same performance as a HMM-based system. The closure–burst phonetic structure of each plosive is well preserved in the speech signal and the boundaries of this unique phonetic structure are very clearly matched. In our experiments we use this natural phonetic segmentation of a plosive and assume that its boundaries are known. We will investigate the possible performance improvement by such an adaptive segmentation in comparison to a fixed segmentation. This is a first step into developing a phonetic knowledge-based detection system which, together with supra-segmental features [3], will enhance the detection results.

The detection of plosives from speech signals is a "tough" problem in phonetics and speech recognition but it is also an important step in many speech applications. E.g. for the coding of speech it might be advantageous to know the position of the plosive sounds and to model them independently as this helps to improve the reconstructed speech quality, see [4]. Some work has been done in designing classifiers that avoid an HMM-based approach. In [5], a detector is considered that marks the time between a closure–burst transition. Another approach is the knowledge-based landmark detector from [6] which is used to distinguish plosive from non-plosive segments. This corresponds to our first experiment.

This report is organized as follows: In Sec. II, we summarize the three experiments which we conduct. Starting from the easiest task, which is to decide whether a plosive is present or not, we differentiate in the second task between voiced plosives and unvoiced plosives. Finally, the last task is the detection of each plosive individually. Sec. III gives a detailed description of the used features which yield the simulation results in Sec. IV. We show that using the additional timing information of the start and end of each phone improves the simulation results significantly.

II. Experiments and Setup

We use the TIMIT corpus [7] for our three experiments because of its meticulous phonetic transcription for each speech file. We use TIMIT as ground truth for each segment to determine the class it belongs to. The three experiments that we conduct are

Exp. 1: Plosive vs. non-plosive classification where closures that belong to the plosive are treated as part of the plosive. Using the TIMIT notation, we try to detect all segments that belong to $\{/b/, /p/, /d/, /t/, /k/, /g/, /q/\}$ and the corresponding closure labels as well.

Exp. 2: A three-way classification: voiced plosives, unvoiced plosives and non-plosives, i.e. we have now three classes: $\{/b/, /d/, /g/$ + closures$\}$, $\{/p/, /t/, /k/$ + closures, $/q/\}$ and the class of all non-plosives.

Exp. 3: Detection of individual plosives, i.e. we have in total seven classes, i.e. six plosive classes $\{/b/\}$, $\{/p/\}$, $\{/d/\}$, $\{/t/\}$, $\{/k/\}$, $\{/g/\}$ and one non-plosive class. This is the most challenging task of all three experiments.

Each experiment is performed twice: In the first run, we segmentize the speech signal in blocks of 10 ms length. For each block, we do a separate classification. The second run uses also a segmentation but the segments are chosen to be identical with the position of each phone as it is annotated in the TIMIT corpus. We compare both runs to evaluate the performance loss if no timing information is used as in the first run. The complete training and test parts of the TIMIT corpus are used for the training and evaluation, respectively.

As classifiers, we use the well-known Bayes and decision tree classifiers [2]. For the Bayes classifier, we assume the feature distribution for each class to be multivariate Gaussian. This classifier is especially suited for our experiments as we have a large number of training and classification patterns, and the Bayes classifier with a multivariate Gaussian distribution is known to be very efficient with respect to its computational complexity. As feature selection algorithm, we use the well-known Sequential Floating Forward Selection (SFFS) algorithm from [8]. However, we modified the SFFS to take the classification rate for each class into account instead of only considering the overall classification rate. This is important as the relative occurrence frequency of the classes differ substantially, e.g. for the detection of plosives vs. non-plosives where the percentage of plosives is relatively small. Without a modification of the SFFS, the classifier would label plosives as non-plosives and by that simple scheme it would achieve a high overall detection rate which is undesirable. Especially the Bayes classifier is prone to that error. Note, that another possibility to deal with the small number of plosives would be a training set regularization where we e.g. randomly choose only as many non-plosives as we have plosives.

III. Features

In this section, we introduce the features that we used for the detection of plosives. All features are based on a 10 ms segmentation. For the case that we use the start and end of a phone to segment the speech signal,we calculate the feature values for all 10 ms segments that fall into the phone interval and then use the average operator to obtain the features.

A. Energy Bands [6]

The first group of features that we use are energy bands [6]. We calculate one energy value for each 10 ms time segment in our experiment. The bands are defined as the frequency intervals 0 Hz $-$ 400 Hz, 800 Hz-1500 Hz, 1200 Hz-2000 Hz, 2000 Hz-3500 Hz, 3500 Hz-5000 Hz, and 5000 Hz-8000 Hz.

B. Energy Envelopes [9]

The next group of features are energy envelopes. Energy envelopes dynamically split the frequency spectrum into bands, depending on the number of bands that should be used. For our experiments, we used four bands which results in the following frequency intervals: 1Hz-8Hz, 8Hz-70Hz, 70Hz-594Hz and 594Hz-5000Hz. This division corresponds to the results given by [9]. Here too, one energy value is calculated for each time segment. Furthermore, we used a lowpass-filtered version of these as additional features.

C. Formant Frequencies and Bandwidths [10]

Another set of features are the formant frequencies and their bandwidth. They are calculated using the LPC approach to obtain an all-pole vocal tract model. Each conjugate complex zero pair corresponds to one formant frequency and its distance to the unit circle gives the bandwith. We use the formulas

$$F_n = \frac{Fs}{2\pi}\left|\text{atan}\left(\frac{\Im\{p_n\}}{\Re\{p_n\}}\right)\right|$$
$$B_n = -\frac{Fs}{\pi}\log(|p_n|)$$

to map a pole p_n of the all-pole model to its frequency F and bandwidth B. The first four formant frequencies and the first four formant bandwidths are used as features for detection of classifiers.

IV. Simulation Results

Exp. 1: The first experiment is plosive vs. non-plosive classification where closures that belong to the plosive are treated as part of the plosive. The results are shown in Table I and II for the case of a fixed and phone-based segmentation. Comparing both tables, we see that the overall classification rate is in the same range for both runs. For the second run, however, the confusion matrix is better balanced between the two classes. The good classification rate for the first run is due to the misclassification of plosives as non-plosives. The reason for this is that we have 236 565 plosive segments opposed to 1 184 951 non-plosive segments. The decision tree classifier provides better results for both cases compared to the

Bayes classifier, although the Bayes classifier is used to select the best features with the SFFS. This shows that the Bayes classifier is not capable of extracting all relevant information that is present in the features.

Fig. 1 on the last page shows the results for the classification rate vs. the number of features when the decision tree classifier was applied for both runs. Clearly, an increasing number of features yields a better classification rate. Table III shows the features that are selected by the SFFS algorithm for ten features. The ordering reflects the time a feature was added to the set, i.e. the first feature was selected first. The best features to distinguish plosives from non-plosives are the energy envelopes and energy bands.

Exp. 2: The second experiment is to differentiate between voiced plosives, unvoiced plosives and non-plosives. Fig. 2 on the last page shows the classification rate vs. the number of features and Table IV and V give the classification rate for a fixed and phone-based segmentation. Similar to Exp. 1, we have a better balanced confusion matrix of about 10% from using the phone-based segmentation.

Table VI shows the best ten features that were selected by the SFFS algorithm for the second experiment. Beside the energy envelopes and energy bands that were selected for Exp. 1, formant frequencies and bandwidths were added to the feature set.

Exp. 3: The third experiment is to differentiate between each plosive ($/p/$, $/t/$, $/k/$, $/b/$, $/d/$, $/g/$) and non-plosives. For this experiment, the Bayes classifier is not considered anymore as it labels all segments as non-plosives and the classification rate for the other classes is therefore zero. The overall classification rate for fixed and phone based segmentation is 79.2% and 73.9%, respectively. The better overall classification rate for the fixed segmentation is, however, due to the misclassification of plosives as non-plosives as can be seen from the confusion matrix in Table VII if compared to Table VIII. The confusion matrix for phone-based segmentation is more balanced than for fixed segmentation and therefore should be preferred. Note, that the classification rates for Exp. 3 are not as good as for the other two experiments. The phone-based segmentation is still better than the fixed segmentation, however, more discriminating features are needed to obtain better results. Fig. 3 on the last page shows the classification rate vs. the number of features.

Table IX shows the best ten features that are selected by the SFFS algorithm for the third experiment. The selected features are similar to the features selected for Exp. 2, but show a different order.

V. Conclusions

In this technical report, we have used a segmental classification approach for the detection of plosives. As this approach cannot by itself take the temporal characteristics of plosives into account, we have

to provide this information by other means. We considered the possibility that the unique boundaries around the closure and burst structure of a plosive are known and we have shown that this additional information about phone boundaries does improve the classification rate significantly. Especially the individual classification rate of the plosive classes is increased. Note, that we used only the mean feature value for each phone. Better classification results are possible by using e.g. the standard deviation or the minimum/maximum value for the phones. So far, we used the labels provided by the TIMIT corpus, but we plan to evaluate our classification architecture using estimated segmentations of the speech signal, e.g. with the help of [11], [12].

Another future direction for our research is to find new features for the classification. One possiblity is the estimation of the voice onset time (VOT) as it has been proven to be helpful for the classification of plosives [13].

REFERENCES

[1] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–285, Feb. 1989.

[2] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 2001.

[3] G. Dogil, *The Pivot model of speech parsing*, Verlag der Österreichischen Akademie der Wissenschaften, Wien, 1988.

[4] T. Unno, T. P. Barnwell, and K. Truong, "An improved mixed excitation linear prediction (MELP) coder," *Proc. IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 245–248, 1999.

[5] P. Niyogi and M. M. Sondhi, "Detecting stop consonants in continuous speech," *The Journal of the Acoustical Society of America*, vol. 111, no. 2, pp. 1063–1076, 2002.

[6] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3417–3430, 1996.

[7] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa timit acoustic phonetic continuous speech corpus," 1993.

[8] P. Pudil, F. J. Ferri, J. Novovicova, and J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion," *Pattern Recognition - Conference B: Computer Vision*, vol. 2, pp. 279 – 283, Oct. 1994.

[9] R. V. Shannon, F. Zeng, K. Kamath, J. Wyngonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, pp. 303–304, 1995.

[10] J. R. Deller, J. H. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, 1999.

[11] L. Golipour and D. O'Shaughnessy, "A new approach for phoneme segmentation of speech signals," in *Interspeech*, 2007, pp. 1933–1936.

[12] G. Flammia, P. Dalsgaard, O. Andersen, and B. Lindberg, "Segment based variable frame rate speech analysis and recognition using a spectral variation function," in *ICSLP*, 1992, pp. 983–986.

[13] P. Niyogi and P. Ramesh, "The voicing feature for stop consonants: recognition experiments with continuously spoken alphabets," *Speech Communication*, vol. 41, pp. 349–367, 2003.

Classifier	# Features	Classification Rate		
		Non-pl.	Pl.	Overall
Decision Tree	1	87.8%	28.9%	78.0%
	2	87.9%	38.4%	79.6%
	3	88.1%	40.5%	80.2%
	10	89.6%	51.0%	83.2%
Bayes Classifier	1	73.5%	56.4%	70.6%
	2	88.8%	48.6%	82.2%
	3	91.9%	35.2%	82.5%
	10	55.5%	92.8%	82.3%

TABLE I: Exp. 1: Classification Rate (Fixed Segmentation)

Classifier	# Features	Classification Rate		
		Non-pl.	Pl.	Overall
Decision Tree	1	82.6%	48.9%	72.8%
	2	84.2%	57.3%	77.0%
	3	85.5%	60.4%	78.8%
	10	89.4%	70.0%	84.2%
Bayes Classifier	1	64.8%	89.7%	71.5%
	2	79.0%	83.2%	80.1%
	3	81.1%	81.8%	81.3%
	10	60.6%	94.3%	69.6%

TABLE II: Exp. 1: Classification Rate (Phoneme-based Segment.)

Feature	Fixed Segm.	Phoneme-based Segm.
1	Low Pass Filtered Third Envelope	
2	First Envelope	
3	Second Envelope	
4	Low Pass Filtered First Envelope	
5	Low Pass Filt. 2nd Env.	Fourth Envelope
6	Low Pass Filt. 4th Env.	Third Envelope
7	Third Envelope	Low Pass Filt. 2nd Env.
8	Sixth Band	Low Pass Filt. 4th Env.
9	Fourth Envelope	First Band
10	Fourth Band	Sixth Band

TABLE III: Exp. 1: Selected Features

Classifier	# Feat.	Classification Rate			
		Non-pl.	Vo. Pl.	Unv. Pl.	Overall
Decision Tree	1	88.9%	4.7%	13.8%	75.9%
	2	88.7%	6.6%	16.5%	76.2%
	3	89.8%	25.2%	32.7%	79.9%
	10	89.2%	27.8%	35.1%	79.8%
Bayes Classifier	1	100.0%	0.0%	0.0%	83.4%
	2	74.3%	0.0%	35.0%	66.0%
	3	60.0%	47.1%	78.0%	61.3%
	10	84.4%	34.8%	32.5%	75.9%

TABLE IV: Exp. 2: Classification Rate (Fixed Segmentation)

Classifier	# Feat.	Classification Rate			
		Non-pl.	Vo. Pl.	Unv. Pl.	Overall
Decision Tree	1	83.8%	12.3%	31.1%	67.7%
	2	83.2%	14.4%	32.7%	67.7%
	3	83.6%	19.0%	33.6%	68.6%
	10	89.4%	41.5%	53.7%	78.5%
Bayes Classifier	1	96.3%	0.0%	18.3%	73.4%
	2	53.2%	62.1%	33.5%	51.0%
	3	65.4%	39.7%	55.9%	61.1%
	10	63.7%	35.0%	86.9%	64.4%

TABLE V: Exp. 2: Classification Rate (Phoneme-based Segmentation)

Feature	Fixed Segm.	Phoneme-based Segm.
1	First Formant	First Envelope
2	First Bandwidth	Second Band
3	Fourth Bandwidth	Second Envelope
4	Third Bandwidth	Low Pass Filt. 2nd Env.
5	Low Pass Filt. 2nd Env.	Third Bandwidth
6	Fourth Formant	Fourth Bandwidth
7	Low Pass Filt. 4th Env.	Third Envelope
8	Second Formant	Low Pass Filt. 3rd Env.
9	Third Formant	Low Pass Filt. 4th Env.
10	Second Bandwidth	Fourth Envelope

TABLE VI: Exp. 2: Selected Features

	Non-pl	/b/	/d/	/g/	/p/	/t/	/k/
Non-pl	**90.3%**	0.5%	1.6%	0.7%	1.1%	3.0%	2.8%
/b/	53.1%	**17.4%**	11.5%	2.8%	5.4%	6.6%	3.2%
/d/	57.8%	3.8%	**18.2%**	3.6%	2.7%	9.0%	4.9%
/g/	58.1%	3.6%	9.0%	**8.3%**	2.5%	6.5%	12.1%
/p/	66.2%	2.5%	3.5%	1.6%	**8.5%**	9.1%	8.6%
/t/	68.5%	1.3%	5.6%	1.9%	3.4%	**12.0%**	7.4%
/k/	63.3%	0.7%	3.1%	3.5%	3.3%	8.1%	**18.0%**

TABLE VII: Exp. 3: Confusion Matrix (10 Features, Decision Tree, Fixed Segm.)

	Non-pl	/b/	/d/	/g/	/p/	/t/	/k/
Non-pl	**89.2%**	1.1%	2.3%	1.0%	1.2%	2.9%	2.4%
/b/	38.9%	**24.6%**	13.1%	4.4%	6.8%	7.7%	4.6%
/d/	40.3%	6.7%	**23.0%**	6.2%	3.6%	14.0%	6.2%
/g/	38.8%	6.2%	13.0%	**15.4%**	2.5%	8.1%	16.1%
/p/	36.7%	6.0%	6.0%	2.1%	**18.7%**	18.2%	12.4%
/t/	42.7%	3.1%	11.3%	2.8%	6.8%	**22.4%**	10.9%
/k/	34.3%	2.4%	5.7%	5.9%	5.5%	12.5%	**33.9%**

TABLE VIII: Exp. 3: Confusion Matrix (10 Features, Decision Tree, Phon.-based Segm.)

Feature	Fixed Segm.	Phoneme-based Segm.
1	Second Formant	
2	Low Pass Filt. 2nd Env.	First Formant
3	Third Formant	Third Bandwidth
4	Fourth Formant	Fourth Bandwidth
5	Fourth Bandwidth	Low Pass Filt. 2nd Env.
6	Third Bandwidth	Second Bandwidth
7	Second Bandwidth	First Bandwidth
8	First Formant	Third Formant
9	First Bandwidth	Fourth Formant
10	Third Envelope	Fourth Band

TABLE IX: Exp. 3: Selected Features

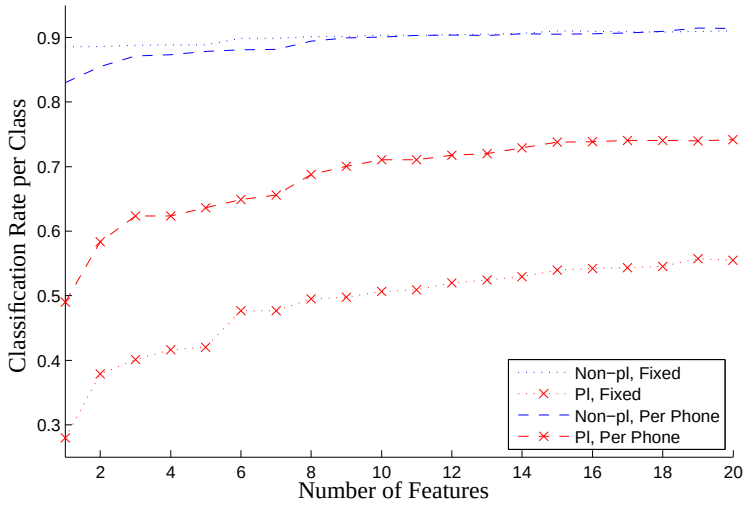

Fig. 1: Exp. 1: Number of Features vs. Classification Rate for Decision Tree Classifier

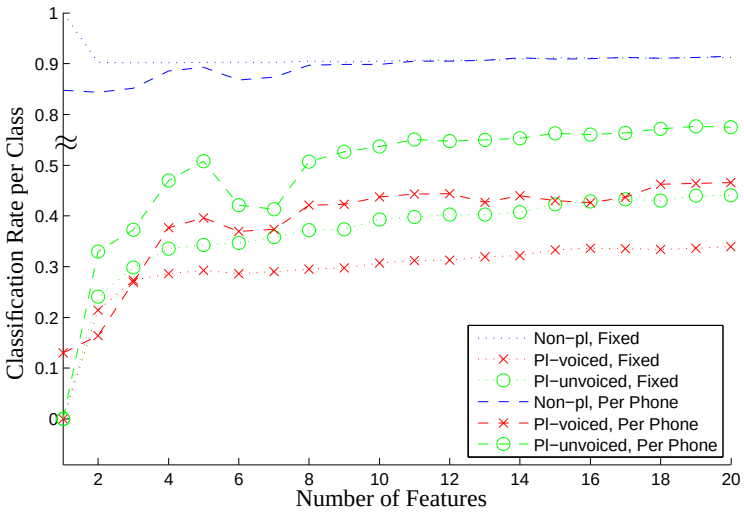

Fig. 2: Exp. 2: Number of Features vs. Classification Rate for Decision Tree Classifier

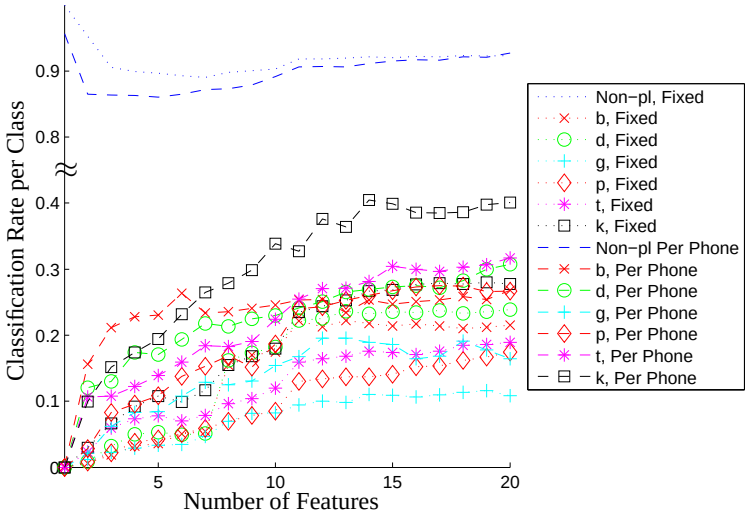

Fig. 3: Exp. 3: Number of Features vs. Classification Rate for Decision Tree Classifier