# Thesis seminar: Unsupervised Segmentation of Continuous Speech Using Vectorautoregressive Modeling

Espoo, Suomi
http://www.acoustics.hut.fi/u/petri/
petri@acoustics.hut.fi

December 14, 2004

# Contents

- Description of the problem
- Signal preprocessing methods
- Description of the proposed algorithm
- Experimental results
- Conclusions and perspectives

# Segmentation

- Segmentation of continuous speech signal into smaller meaningful units (phonemes, syllables)
- Articulatory movements produce changes in acoustic signal
- Proposed method based on detecting these rapid changes in speech spectrum using only the speech signal
- No additional information given for the system
- No training
- Prediction of speech spectrum using vectorautoregression
- Forward and backward prediction used
- Error increases at segment boundaries

# Vector Autoregressive Model

- $VAR(p)$ model is defined as
- $\mathbf{y}_t = \mathbf{A}(1)\mathbf{y}_{t-1} + \ldots + \mathbf{A}(p)\mathbf{y}_{t-p} + \mathbf{v} + \mathbf{u}_t$
- $\mathbf{A}(i)$ are fixed $(K \times K)$ matrices, $\mathbf{v}$ is $(K \times 1)$ vector allowing non-zero mean, $\mathbf{u}_t$ vector of white noise
- Parameters are estimated from time series using multivariate least squares estimation
- VAR(1) model predicts the vector at time $t$ from the vector at time $t-1$
- $\hat{\mathbf{y}}_t = \mathbf{A}\mathbf{y}_{t-1} + \mathbf{v}$
- Model $\mathbf{A}$ estimated from multivariate time-series using least squares estimation. Error is the one-step prediction error between subsequent vectors within the data window

# Proposed Algorithm

- Digital speech signal $s(n)$ converted to short-time features $\mathbf{y}_t$ each being $(p \times 1)$ vector
- Define $\mathbf{A}_t$ as the VAR(1) model computed from the $L$ data vectors ending at vector at time $t$:

$$\mathbf{A}_t = VAR_{LSE}(\mathbf{y}_{t-L+1} \ldots \mathbf{y}_t)$$

- $L$ should correspond to average length of a steady state of a phoneme in speech

- For each vector $\mathbf{y}_t$ compute recursively $M$ estimates with models $\mathbf{A}_{t-M} \ldots \mathbf{A}_{t-1}$

$$\begin{aligned} \hat{\mathbf{y}}_{t1} &= \mathbf{A}_{t-1}\mathbf{y}_{t-1} \\ \hat{\mathbf{y}}_{t2} &= \mathbf{A}_{t-2}^{2}\mathbf{y}_{t-2} \\ &\vdots \\ \hat{\mathbf{y}}_{tM} &= \mathbf{A}_{t-M}^{M}\mathbf{y}_{t-M} \end{aligned}$$

- From these we get relative errors

$$\begin{aligned} e_{t1} &= \frac{(\mathbf{y}_t - \hat{\mathbf{y}}_{t1})^T(\mathbf{y}_t - \hat{\mathbf{y}}_{t1})}{\mathbf{v}_t^T \cdot \mathbf{v}_t} \\ e_{t2} &= \frac{(\mathbf{y}_t - \hat{\mathbf{y}}_{t2})^T(\mathbf{y}_t - \hat{\mathbf{y}}_{t2})}{\mathbf{v}_t^T \cdot \mathbf{v}_t} \\ &\vdots \\ e_{tM} &= \frac{(\mathbf{y}_t - \hat{\mathbf{y}}_{tM})^T(\mathbf{y}_t - \hat{\mathbf{y}}_{tM})}{\mathbf{v}_t^T \cdot \mathbf{v}_t} \end{aligned}$$

- The median value of these errors will represent the final error at time $t$

$$e_t = \mathrm{median}(e_{t1} \ldots e_{tM})$$

- The small values are emphasized by taking the logarithm

$$E_t = 10 \log_{10}(1 + e_t)$$

- So far the model has been used to predict the values of the multivariate time-series for the future values of $\mathbf{y}$. The model can also be used to estimate values for the vectors before the model data-window
- Time reverse the original signal and perform the same $VAR$ analysis

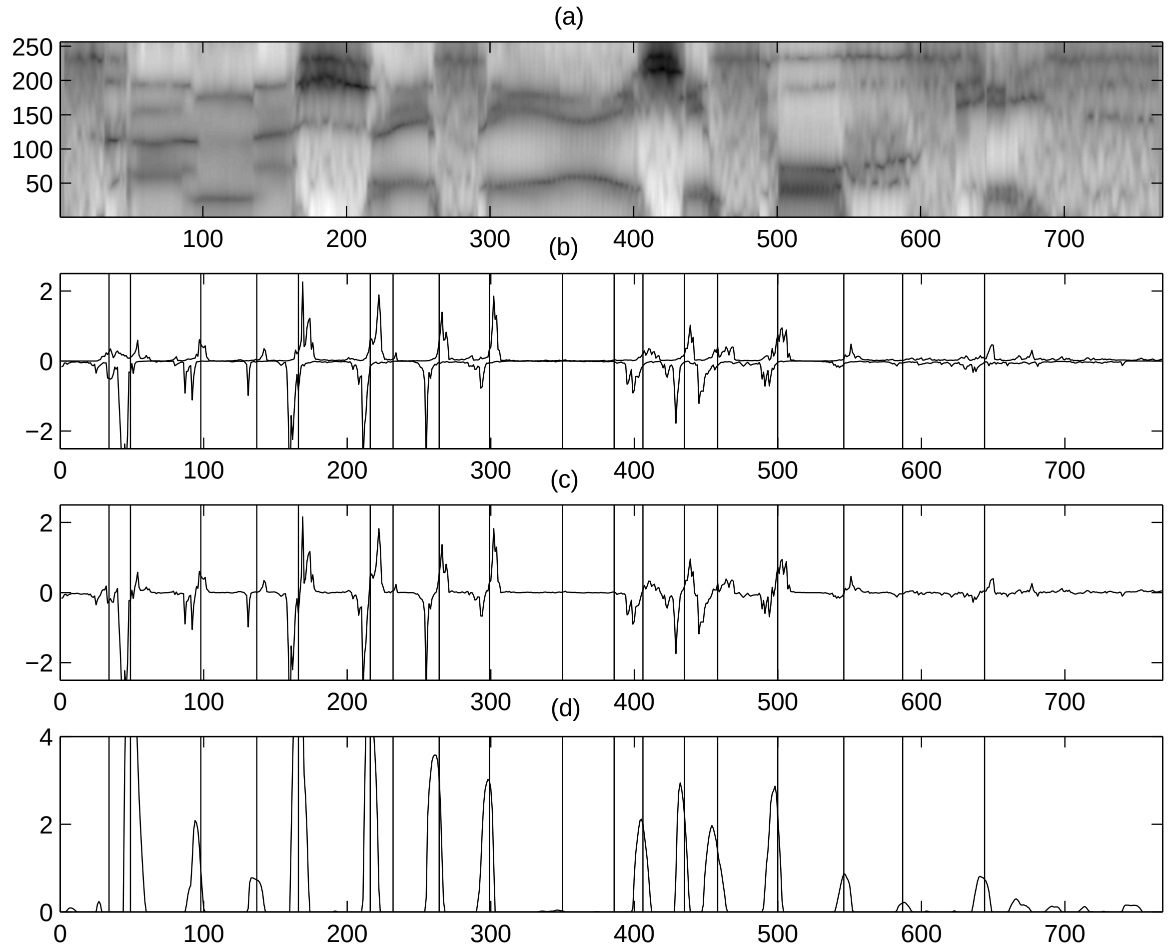
- Let us denote the error signals obtained this way with $E_{t+}$ and $E_{t-}$
- Errors are combined

$$E_{t*} = E_{t+} - E_{t-}$$

- Resultant error $E_{t*}$ should have a large negative peak before rapid change in the signal and large positive peak after the change
- To help the detection of these points the signal is filtered with $h(t)$

$$h(t) = \begin{cases} \frac{t}{d} + 1 & -d < t < 0 \\ 0 & t = 0 \\ \frac{t}{d} - 1 & 0 < t < d \end{cases}$$

- The value of $d$ set to match the width of the peaks in $E_{t*}$


(a)
250
200
150
100
50
100
200
300
400
500
600
700
(b)
2
0
−2
(c)
(d)
4

# Experimental Results: Data

- Method was tested on Finnish
- 3 speakers used (one female, two males)
- 201 sentences read (some really artificial sentences)
- 20.05 kHz sampling frequency
- 14th order frequency warped line spectrum pairs computed every 3ms used as the time-series $\mathbf{y}$

# Experimental Results: Evaluation Criterion

- Evaluation of performance is not straightforward
- Automatic segmentation compared with manual segmentation
- Types of error: *insertion*, *deletion*
- Precision/correctness $C$ describes the portion of segment boundaries placed correctly

$$C = \frac{HITS}{HITS+INSERTIONS}$$

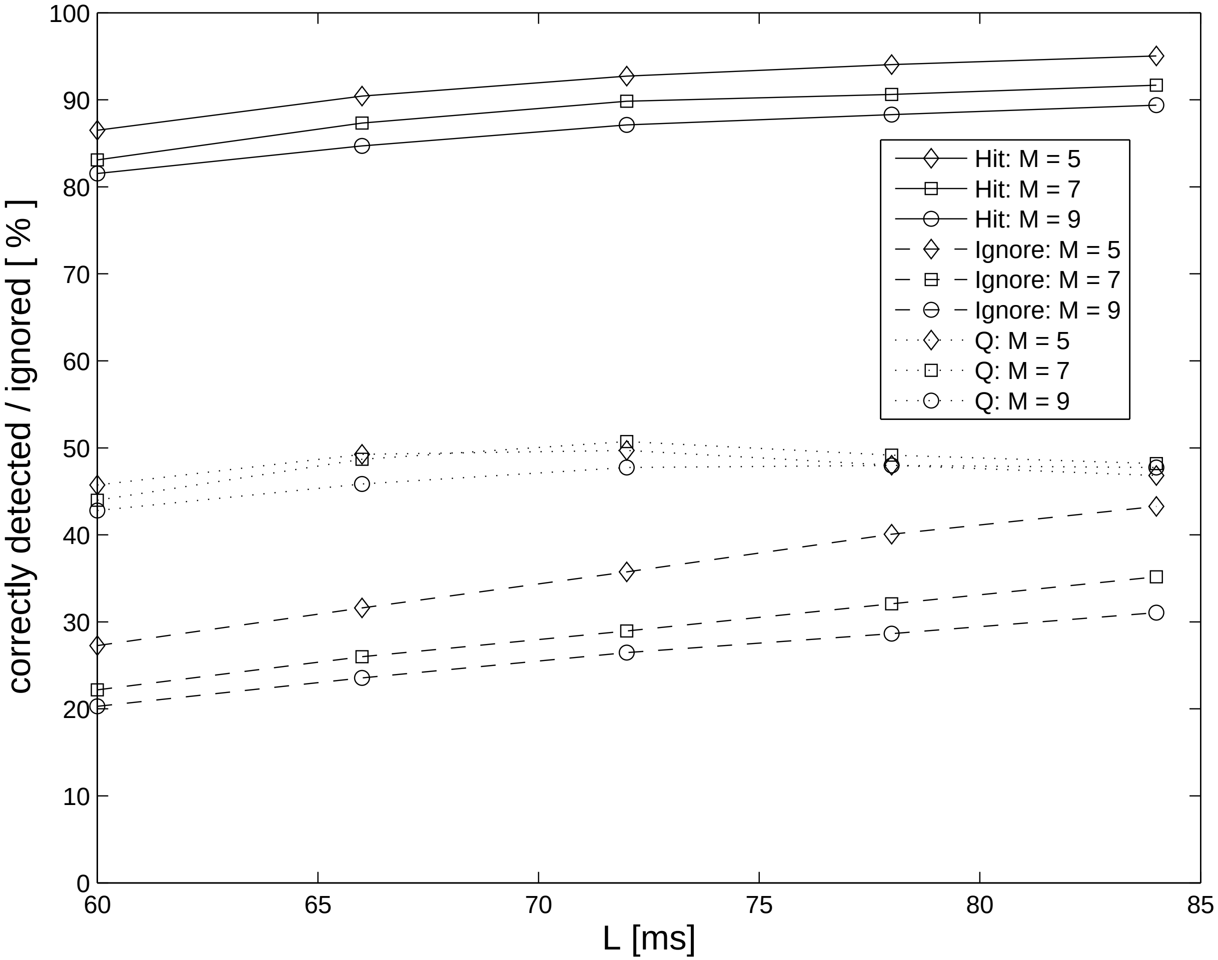
- Quality

$$Q = \frac{HITS-DELETIONS-INSERTIONS}{ALL}$$

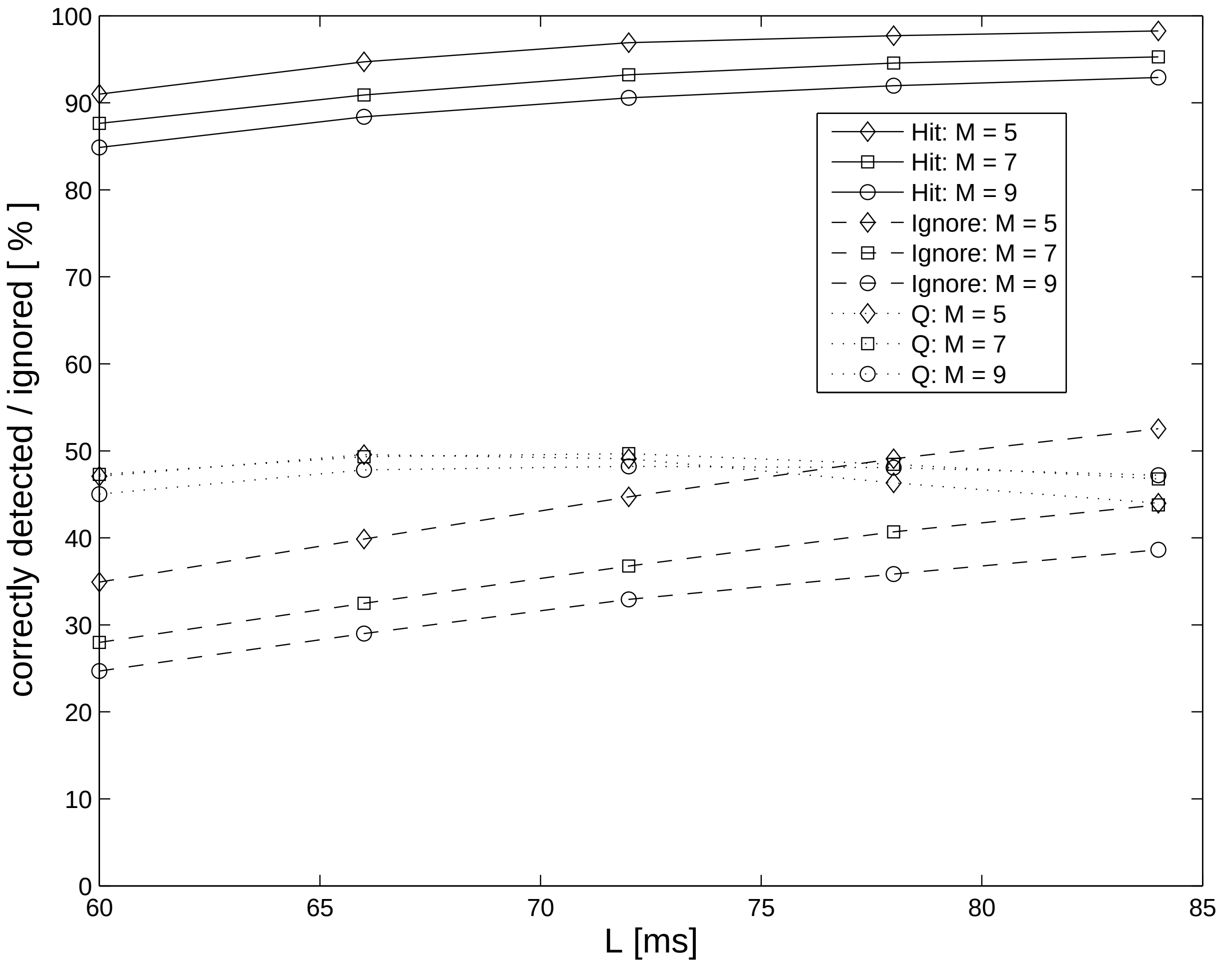
# Experimental Results: Overall Results

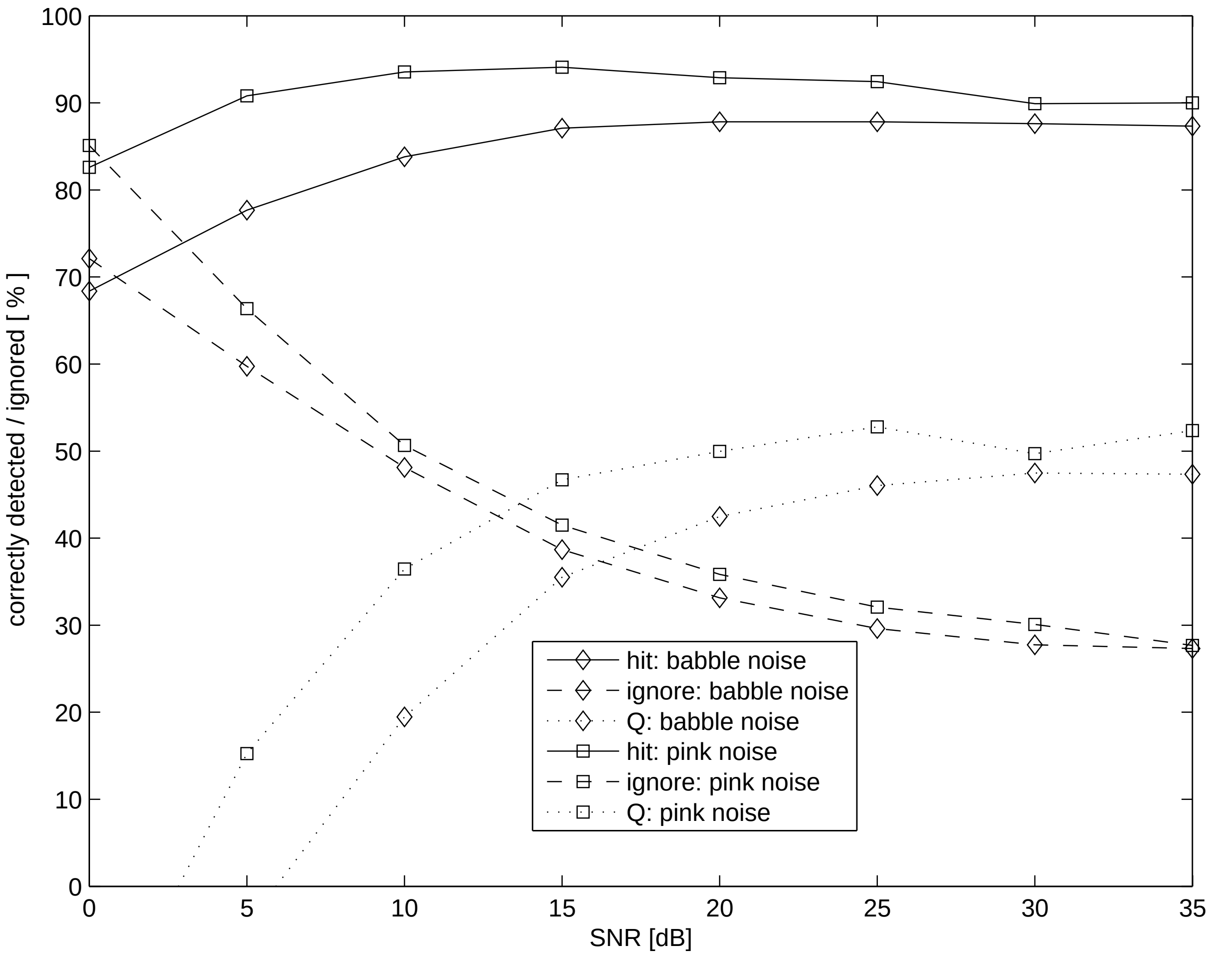
Table 1: Segmentation results for three different speakers (C: Correctness, D: Deletions, Q: Quality), $M = 7$, $L = 66ms$, threshold $= 0.2$

| | C | D | Q |
|---|---|---|---|
| Male 1 | 87.3% | 26.0% | 48.7 |
| Male 2 | 88.2% | 32.6% | 43.7 |
| Female | 91.5% | 35.2% | 47.8 |



correctly detected / ignored [ % ]
L [ms]
Hit: M = 5
Hit: M = 7
Hit: M = 9
Ignore: M = 5
Ignore: M = 7
Ignore: M = 9
Q: M = 5
Q: M = 7
Q: M = 9


100
90
80
70
60
50
40
30
20
10
0
60
65
70
75
80
85
L [ms]
correctly detected / ignored [ % ]
Hit: M = 5
Hit: M = 7
Hit: M = 9
Ignore: M = 5
Ignore: M = 7
Ignore: M = 9
Q: M = 5
Q: M = 7
Q: M = 9


correctly detected / ignored [ % ]
SNR [dB]
100
90
80
70
60
50
40
30
20
10
0
0
5
10
15
20
25
30
35
hit: babble noise
ignore: babble noise
Q: babble noise
hit: pink noise
ignore: pink noise
Q: pink noise

# Other Analysis

- In this master's thesis following things were also investigated:
  - Errors in terms of phoneme classes
  - Emount of $E_*$ at segment boundaries
  - Temporal deviations from manually asigned segment boundaries
  - (Computational load)

# Conclusion

- Method to detect unpredictable auditory time-frequency changes in acoustic signals was presented
- Based on VAR-modeling of multivariate time-series
- Fully unsupervised. Does not use any *a priori* knowledge of the signals chosen for segmentation
- Method was tested on Finnish
- The results show that the method works for both male and female speakers
- The segmentation is reliable between classes that produce abrupt change at segment boundary
- Vowel-vowel pair is the most difficult to detect

- Future work:
  - Testing other representations for the signal (energy, zero-crossing rate ...)
  - Different strategies for selection of segment boundaries from $E_*$
  - Different speech material