IDIAP RESEARCH REPORT

NEUROMORPHIC BASED OSCILLATORY DEVICE FOR INCREMENTAL SYLLABLE BOUNDARY DETECTION

Alexandre Hyafil Milos Cernak

Idiap-RR-14-2015

JUNE 2015

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Neuromorphic Based Oscillatory Device for Incremental Syllable Boundary Detection

Alexandre Hyafil, Milos Cernak

June 10, 2015

Abstract

Syllables are considered as basic supra-segmental units, used mainly in prosodic modelling. It has long been thought that efficient syllabification algorithms may also provide valuable cues for improved segmental (acoustic) modelling. However, the best current syllabification methods work offline, considering the power envelope of whole utterance.

In this paper we introduce a new method for detection of syllable boundaries based on a model of speech parsing into syllables by neural oscillations in human auditory cortex. Neural oscillations automatically lock to speech slow fluctuations that convey the syllabic rhythm. Similarly as humans encode speech incrementally, i.e., not considering future temporal context, the proposed method works incrementally as well. In addition, it is highly robust to noise. Syllabification performance for English and different noise conditions was compared to the existing Mermelstein and group delay algorithms. While the performance of the existing methods depend on the type of noise and signal to noise ratio, the performance of the proposed method is constant under all noise conditions.

Index Terms: speech recognition, syllable identification, neuromorphic systems

1 Introduction

Although automatic speech recognition (ASR) systems generally disregard syllables as valuable representations, there is a general agreement that syllables provide a stable construct across languages that does not suffer boundary indetermination problems as phonemes do (Greenberg, 1998). Detecting the timing of syllable boundaries may thus provide additional information that could improve the performance of phoneme-based ASR systems, and could be employed in a variety of automatic speech applications. Syllables also play a crucial role in prosodic analysis and synthesis, and robust detection of syllables boundaries is often required (Cernak et al., 2013).

A syllable is structurally divisible into three parts, the onset, nucleus and coda. Greenberg (1998) found that syllabic onsets are generally preserved in spoken utterance, while nuclei and codas are more often deleted. Increases in speaking rate result also in more deletions and mutations of most phonetic phonetic constituents as syllabic onsets (Greenberg, 1996). Therefore onset time information can be considered as more robust information compared to syllable nuclei timing, for example such as proposed by de Jong and Wempe (2009). In this work we focus on syllable onset detection, called further also syllable boundary detection.

By contrast to conventional syllabification models that work offline, humans "encode" speech in an incremental fashion, i.e., encoded speech does not depend on future temporal context (similar to causality in digital signal processing theory) (Levelt, 1993). We are therefore interested in an incremental syllabification method that can be directly applied to incremental speech processing systems such as in Cernak et al. (2015). We hypothesise that a biologically plausible method would fulfill this requirement.

Recent evidence from psychoacoustics and neuroimaging studies indicate that in humans, the syllabification process is performed by slow neural oscillations (3-8 Hz) in auditory cortex that track fluctuations in speech power of similar time scale (Giraud and Poeppel, 2012). A computational model of self-generated neural oscillations showed as a proof-of-concept that: (i) such neural oscillations can reliably signal syllable boundaries; (ii) detected syllable boundaries can improve recognition of linguistic units in a parallel neural pathway (Hyafil et al., 2012). In such model, coupled excitatory and inhibitory neurons intrinsically synchronize around 6 Hz, and automatically lock to edges in speech amplitude that convey the syllabic flow.

In this work we investigate whether the neural model could be adapted into an efficient ready-to-use syllabification algorithm. The original neural model (Hyafil et al., 2012) used auditory channels, i.e. the output from a model mimicking all precortical treatment of acoustic signals in the human brain. The model included normalising speech signal over a long segment (e.g. a sentence), decomposition into 32 frequency bands ('channels') through IIR convolution, lateral inhibition between neighbouring channels, half-wave rectification, leaky integration and finally downsampling to a time step of 10 ms (Chi et al., 2005). To simplify feature calculation process, we propose an extended neural model that works with conventional Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) coefficients as the input speech feature representation.

Second, a resilient syllable detector should be able to perform efficiently even in the presence of moderate noise. To our knowledge, previous algorithms have only been evaluated under noiseless conditions. We thus evaluated syllable boundary detection under noisy conditions in our algorithm and in two classic syllabification algorithms: the Mermelstein (Mermelstein, 1975) and group delay (Kamakshi Prasad et al., 2004) algorithms.

The paper is structured in the following way. Section 2 introduces oscillation based detection including the parameters optimization procedure. Section 3 describes performance evaluation of the proposed method and of the two alternative algorithms. Section 4 presents the results, and finally Section 5 concludes the paper with discussion and and an outline of future work.

2 Oscillation based detection of syllable boundaries

2.1 Intrinsic oscillatory mechanism

The detector is based on an interconnected network of leaky integrate-and-fire (LIF) neurons. Its principles are based on findings on the role of slow neural oscillations in auditory cortex for natural speech parsing (Giraud and Poeppel, 2012; Hyafil et al., 2012). In essence, the network is composed of $n_E = 10$ excitatory and $n_I = 10$ inhibitory neurons. Synchronization occurs in the network through a burst of inhibitory spikes occurring after receiving sufficient excitatory input. The dynamics of each neuron membrane potential V_i and synaptic activation variable s_i follows :

$$V_i(t+1) = V_i(t) + \frac{I_i(t)}{C}dt, \tag{1}$$

where C stands for membrane currents and $I_i(t)$ stands for membrane currents that consists of:

$$\begin{aligned} I_i(t) =& I_i^{leak}(t) + I_i^{DC} + \\ & k_i I_i^{ext}(t) + I_i^{syn}(t) + \eta_i(t), \end{aligned} \tag{2}$$

where the partial currents are the leak current I_i^{leak}, a constant current I_i^{DC}, an external (speech) current I^{ext} (only for excitatory neurons), a synaptic current $I_i^{syn}(t)$ and a noise current consisting of i.i.d. gaussian noise $\eta_i(t)$ of variance $\sigma_i(t)$. Voluntarily adding noise to an automatic system is quite uncommon, but here such noise adds flexibility to the oscillatory network, allowing it to rapidly lock to speech input. Whenever membrane potential reaches threshold V_i^{thr}, the neuron emits a spike that is propagated in the network and V_i is reset to V_i^{res}. Leak currents follow:

$$I_i^{leak}(t) = g^L(V_i^L - V_i(t)) \tag{3}$$

Synaptic currents follow:

$$I_i^{syn}(t) = \sum_j s_{ij}(t)(V_j^{syn} - V_i(t)) \tag{4}$$

where j stands for each neuron (either excitatory or inhibitory) connecting to neuron i and $s_{ij}(t)$ is the activation variable for the j-to-i synapse. The dynamics of this variable follows:

$$s_{ij}(t+1) = s_{ij}(t) + \frac{r_{ij}(t) - s_{ij}(t)}{\tau_i^D}dt \tag{5}$$

$$r_{ij}(t+1) = r_{ij}(t) + \delta(spk_j(t))g_{ij} - \frac{r_{ij}(t)}{\tau_i^R}dt \tag{6}$$

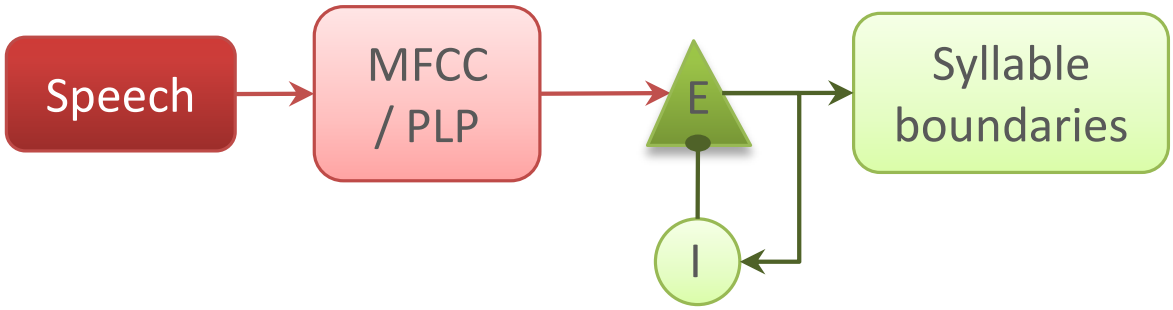


Figure 1: *Schematic diagram of oscillation-based syllable boundary detection algorithm. E represent excitatory neurons, and I inhibitory neurons.*

where $\delta(spk_j(t))$ is 1 if neuron j emits a spike at time t, 0 otherwise, τ_i^R and τ_i^D are respectively the rising and decay time of synaptic activation.

For both types of speech representations, MFCCs and PLP coefficients, we reduced the multidimensional signal (n=13) to a single temporal signal by some spectral weight w_{sp} and then convolved the signal with a temporal kernel k_{temp} spanning 4 frames:

$$I_i^{ext}(t) = k_i \lambda_{temp} * (\sum_{ch} w_{sp} X) \tag{7}$$

where X is the matrix of PLP/MFCC coefficients.

A putative syllable boundary was declared for each inhibitory spike burst, that is whenever there were at least 2 inhibitory spikes occurring within a window of 15 ms.

The neural oscillation keeps running even in the absence of acoustic input and thus provides putative syllable boundaries even for silence periods. To prevent this we masked the putative boundaries using a silence detection.

2.2 Parameters optimization

Parameter fitting was performed using a subsample of 1000 noiseless sentences from the TIMIT corpus (all 10 sentences for speakers indexed 1-100) (Consortium, 1993). We used syllabification program tsylb2 (Fisher, 1996) to convert phonetician-labelled phonemes and phoneme boundaries into syllables and syllable boundaries.

We first determined spectral weights and temporal kernel by finding values such that the weighted and convolved signal $y(t) = \lambda_{temp} * (w_{sp} X)$ (where X is the pre-whitened matrix of speech features) maximized its averaged value at time of syllable boundaries $< y(t_{boundaries}) >$. Then we only retained the spectral weights and computed a refined value of the temporal kernel by using a simplified single neuron model called GLM point process model for spike trains[1]. By providing the weighted signal $w_{sp}X$ as input and syllable boundaries as target output, the algorithm optimizes temporal kernel so that the single neuron output resembles the target output as closely as possible .

Most network parameters were simply the same as those used in the original modelling work (Hyafil et al., 2012), which were specifically optimized for auditory channels (see original publication for parameter values). The only exceptions are parameters k_E and I_E^{dc} (index E stands for excitatory neurons). These parameters were optimized separately for PLP and MFCC inputs, by performing a parameter search minimizing the syllabic distance over the 1000 sentences of the training set (values in Tables 1 and 2).

[1] http://pillowlab.cps.utexas.edu/code_GLM.html

Table 1: *MFCC algorithm parameter values*

parameter	value
I_E^{DC}	2.499
k_E	0.0015

Table 2: *PLP algorithm parameter values*

parameter	value
I_E^{DC}	0.6736
k_E	-5.582

3 Performance evaluation

Performance was tested on a distinct subset of 3620 sentences from the TIMIT corpus under clean speech and noisy conditions. The testing set was constructed from the speakers indexed $101 - 462$ and the sentences indexed $1 - 10$. We applied additive noise with SNR ranging from -20 to 20 dB to all test sentences. We used the RSG-10 (Steeneken and Geursten, 1988) collection as the source of noise. We selected three types of the RSG-10 noises:

1. white noise: acquired by sampling high-quality analog noise generator (Wandel & Goltermann) that exhibits equal energy per Hz bandwidth,
2. pink noise: acquired by the same noise generator exhibiting equal energy per 1/3 octave,
3. babble noise: acquired by recording samples from 1/2" B&K condensor microphone. The source of this babble is 100 people speaking in a canteen. The room radius was over two meters; therefore, individual voices are slightly audible.

To add noise, we used Guenter Hirsch's FaNT tool[2], using the "-m snr 8khz" option that computes an unweighted, fullband SNR.

Hits and false alarm rates have previously been used to evaluate syllabification performance (Villing et al., 2006), but a combined signal detection measure such as d' could not be used since correct rejections cannot be defined. Instead we used a distance measure between point process realizations (here syllabic boundary times) that was originally introduced to measure distance between spike trains (Victor and Purpura, 1997). We used 50 ms as the shift cost, i.e. the maximal time discrepancy between an actual boundary and its corresponding predicted boundary. The overall score was the summed of syllabic distance over a corpus, normalized by the sum of the number of overall predicted and actual boundaries in the corpus.

3.1 Alternative algorithms

Performance was compared with two existing algorithms for syllabification: the Mermelsteing algorithm (Mermelstein, 1975) and the group delay algorithm (Nagarajan et al., 2003; Kamakshi Prasad et al., 2004). Both algorithms identify syllable boundaries as local minima in speech power/envelope. Specifically, the Mermelstein algorithm looks for local maxima of the difference between speech envelope and its convex hull. The group delay algorithm looks for positive peaks of a so-called group delay function that is computed from the short term energy of the speech signal. We used the latest group delay implementation downloaded from IIT Madras[3] with default parameters. It also outputs doubtful syllabic segments that we did not consider in the evaluation. We also used as a control model a purely rhythmic signal that outputs putative boundaries regularly at a 7 Hz irrespective of the speech input, thus constituting a chance level reference (the rate was optimized over the training data set).

To eliminate impact of difficult silence detection with highly noisy data, we performed a Praat (Boersma, 2001) silence detection on noise-clean recordings with a -36 dB silence detection threshold and 100 ms minimal silent and sounding intervals. This silence removal enabled masking of the putative boundaries produced by our neural oscillator model, and optimising of group delay parameters

[2] http://dnt.kr.hs-niederrhein.de/download/fant.tar.gz

[3] http://lantana.tenet.res.in/website_files/resources/Syllable_segmentation.tar

as it internally includes silence detection. Thus we employed the same silence removal procedure for all three algorithms.

4 Results

The syllabification process with neural oscillator is illustrated in Figure 2: inhibitory bursts closely matched actual syllable boundaries for that sentence.

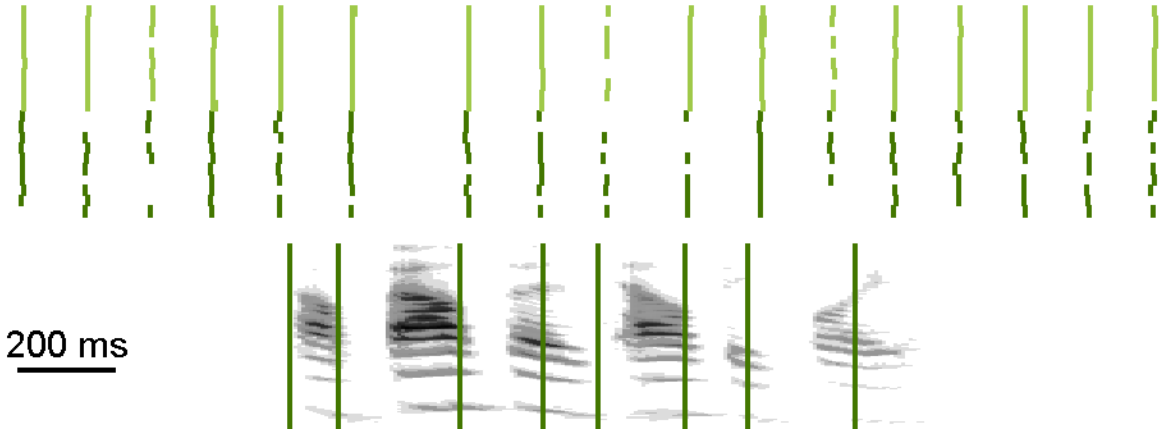


Figure 2: *Model output for one examplar sentence ('Alfafa is healthy for you'). Dark green ticks on top represent excitatory neurons spikes, light green ticks represent inhibitory neuron spikes. Vertical lines on top of spectrogram represent actual syllable boundaries.*

Performance for the distinct algorithms over the full test dataset is similar for distinct types of noise (Figure 3). Mermelstein and neural oscillator with PLP input had comparable levels for low noise conditions (SNR $>$ 10 dB). Performance for the neural oscillator with PLP input was remarkably maintained in moderate and high noise conditions. By contrast, performance for the Mermelstein algorithm was intriguingly found to increase for moderate noise conditions (SNR between -5 and 5 dB), and to severely deteriorate for high level conditions, performing worse than the rhythmic control model. Such deterioration occurred earlier for babble and pink noise than for white noise. Syllabic boundary detection using neural oscillation with MFCC was poor, yet very resilient to high levels of noise as for the PLP input. Performance for group delay followed a similar trend to the Mermelstein algorithm: high performance for low to moderate noise with severe deterioration in high noise environments.

Figure 4 shows modified ROC with babble noise. As correct rejections could not be defined, we instead computed False Alarm Rate (FAR) as $\mathrm{FPR} = \mathrm{FA}/(\mathrm{FA} + \mathrm{TP})$, where FA is the number of false alarms and TP the number of true positives. It can be seen than the group delay is the most conservative of the algorithms with fewer hits and false alarms than other algorithms. Noise increases the incidence of false alarms and moderately affects hit rate.

5 Discussion

We have presented a biologically plausible method of syllable boundaries that (i) works incrementally and (ii) is robust to highly noisy speech (SNR < -5 dB). While the performance of the existing methods depend on the type of noise and signal to noise ratios, the performance of the proposed method is constant under all noise conditions.

The neural oscillation algorithm provided robust incremental prediction for syllable boundaries using PLPs as speech input. Performance was resilient to very high level of background noise, for all types of noises. We expect the algorithm output could be used in a variety of speech applications, from unsupervised speech data labelling to ASR and low bit rate speech coding devices.

The neural oscillator based syllable boundary detector is implemented in Matlab (the parameter optimisation routines). The running executable is implemented in C and the code is available as open-source code at the following address: https://github.com/ahyafil/sylb_boundary.

6 Acknowledgements

We thank A-L Giraud for fruitful discussions at the origin of this project.

References

Paul Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2001.

Milos Cernak, Xingyu Na, and Philip N. Garner. Syllable-Based Pitch Encoding for Low Bit Rate Speech Coding with Recognition/Synthesis Architecture. In *Proc. of Interspeech*, pages 3449–3452, August 2013. URL http://publications.idiap.ch/index.php/publications/show/2602.

Milos Cernak, Philip N. Garner, Alexandros Lazaridis, Petr Motlicek, and Xingyu Na. Incremental Syllable-Context Phonetic Vocoding. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 23(6):1019–1030, June 2015. doi: 10.1109/TASLP.2015.2418577.

Taishih Chi, Powen Ru, and Shihab a Shamma. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906, August 2005. ISSN 0001-4966.

Linguistic Data Consortium. TIMIT Acoustic-Phonetic Continuous Speech Corpus, 1993.

NivjaH de Jong and Ton Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2):385–390, 2009. doi: 10.3758/brm.41.2.385. URL http://dx.doi.org/10.3758/brm.41.2.385.

W. M. Fisher. tsylb2, 1996. URL http://www.nist.gov/speech/tools.

Anne-Lise Giraud and David Poeppel. Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, 15(4):511–517, March 2012. ISSN 1097-6256. doi: 10.1038/nn.3063.

Steven Greenberg. Understanding speech understanding: Towards a unified theory of speech perception. In *Proc. of the ESCA Workshop on the "Auditory Basis of Speech Perception"*, pages 1–8, 1996.

Steven Greenberg. Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation. In *Proc. of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 47–56, 1998.

Alexandre Hyafil, Lorenzo Fontolan, Claire Kabdebon, Boris Gutkin, and Anne-Lise Giraud. Speech encoding by coupled cortical theta and gamma oscillations. *eLife*, 15:in revision, 2012.

V. Kamakshi Prasad, T. Nagarajan, and Hema A. Murthy. Automatic segmentation of continuous speech using minimum phase group delay functions. *Speech Communication*, 42(3-4):429–446, April 2004. ISSN 01676393. doi: 10.1016/j.specom.2003.12.002. URL http://dx.doi.org/10.1016/j.specom.2003.12.002.

Willem J. M. Levelt. *Speaking: From Intention to Articulation (ACL-MIT Series in Natural Language Processing)*. A Bradford Book, August 1993. ISBN 0262620898.

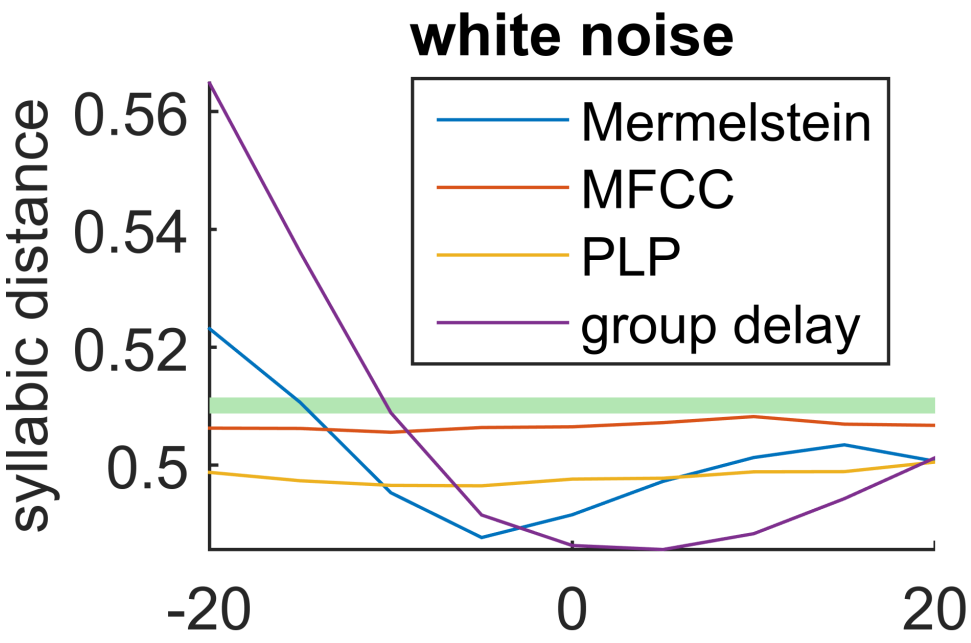
P. Mermelstein. Automatic segmentation of speech into syllabic units. *J. Acoust. Soc. Am*, 58(4):880–883, 1975.

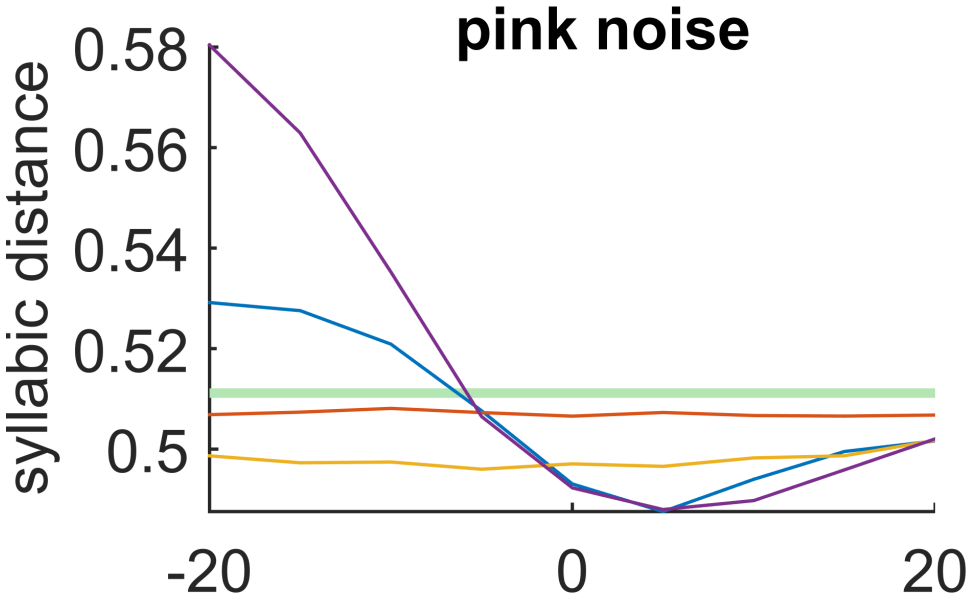
T Nagarajan, HA Murthy, and RM Hegde. Segmentation of speech into syllable-like units. *Energy*, pages 2893–2896, 2003.

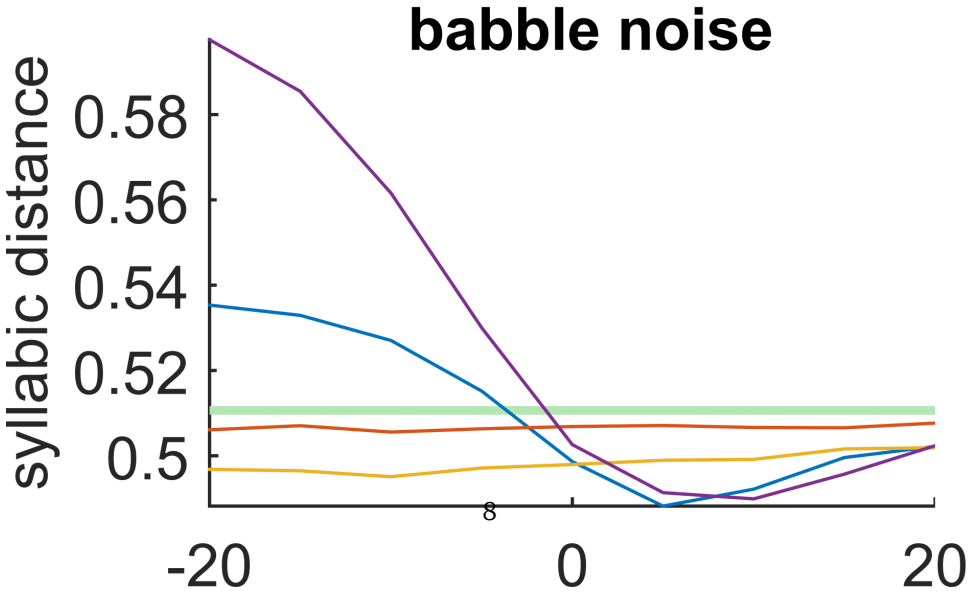
F. Steeneken and F. Geursten. Description of the RSG-10 noise database. Technical report, TNO Institute for Perception, The Netherlands, 1988.

Jonathan Victor and Keith Purpura. Metric-space analysis of spike trains: theory, algorithms and application. *Network: Computation in Neural Systems*, 8(2):127–164, May 1997. ISSN 0954-898X. doi: 10.1088/0954-898X/8/2/003.

Rudi Villing, Tomas Ward, and Joseph Timoney. Performance limits for envelope based automatic syllable segmentation. *IET Irish Signals and Systems Conference (ISSC 2006)*, pages 521–526, 2006. doi: 10.1049/cp:20060489.


white noise
Mermelstein
MFCC
PLP
group delay
syllabic distance
0.56
0.54
0.52
0.5
-20
0
20


pink noise
syllabic distance
0.58
0.56
0.54
0.52
0.5
-20
0
20


babble noise
syllabic distance
0.58
0.56
0.54
0.52
0.5
-20
0
20

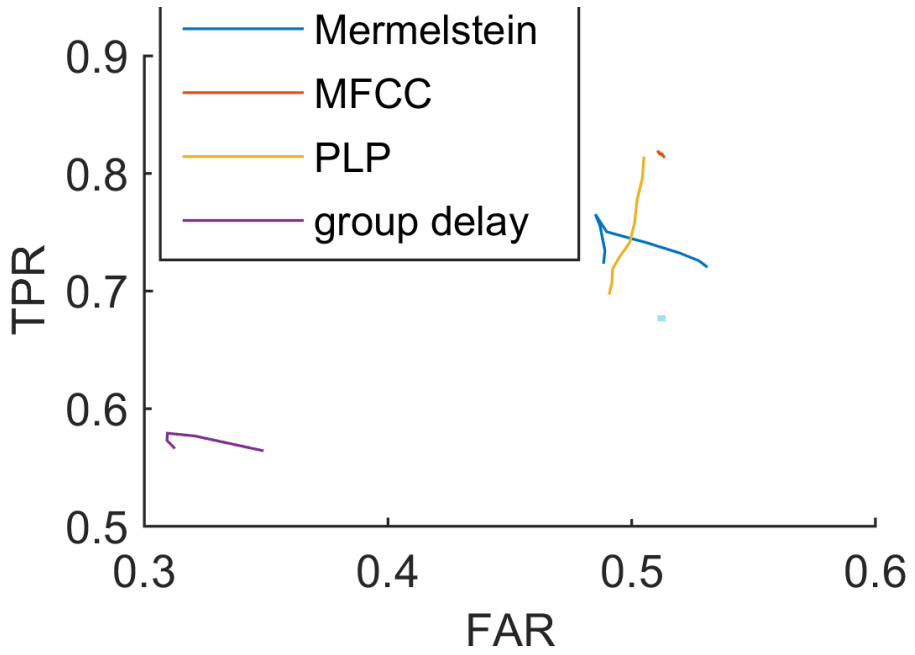


Figure 4: *Modified ROC with babble noise. The x-axis represents the False Alarms Rate (FAR) – false alarms, and y-axis represents the True Positive Rate (TPR) – correct hits of the syllable boundaries. Circles with darker fillings indicate values for higher level of noise. SNR range from -20 dB (black filling) to +20 dB (black filling).*