

Voice Activity Detection

П. А. Холявин

p.kholyavin@spbu.ru





Определение пауз (Praat)

1. Вычисление контура интенсивности
2. Интервалы над и под заданным порогом отмечаются как речь и пауза соответственно
3. Короткие звучащие интервалы убираются
4. Короткие паузы убираются

По умолчанию:

порог = -25 dB

минимальный звучащий интервал = 0.1 с

минимальная пауза = 0.1 с



Voice Activity Detection

1. Извлечение признаков

2. Принятие решения

А какие могут быть признаки?..



Voice Activity Detection

1. Спектральная мощность
2. Соотношение сигнал/шум (SNR)
3. Zero-crossing rate
4. Автокорреляция
5. Спектральная энтропия
6. Формантная структура
7. Стационарность
8. Темпоральная структура (периодические изменения энергии)



Voice Activity Detection (Praat)

Y. Ma & A. Nishihara (2013): "Efficient voice activity detection algorithm using long-term spectral flatness measure.", EURASIP Journal on Audio, Speech, and Music Processing, 2013:21

Long-term spectral flatness measure (LSFM)

LSFM высокий → спектр однородный (шум)

LSFM низкий → спектр неоднородный



Voice Activity Detection (Praat)

N_w, N_{sh} – длина окна и шаг (в отсчётах)

R – количество окон для анализа (“длинное” окно)

M – количество окон для вычисления мгновенного спектра

$X(p, f_k)$ – значение спектра Фурье в окне p и частоте f_k (с окном Ханна)

Обрабатываемые частоты – от 500 до 4000 Гц

(вопрос: как, имея ЧД и размер окна ДПФ, определить номер отсчёта для частоты 500 Гц?)

1. Спектр по методу Уэлча-Бартлетта:

$$S(n, f_k) = \frac{1}{M} \sum_{p=n-M+1}^n |X(p, f_k)|^2$$



Voice Activity Detection (Praat)

2. Среднее геометрическое по всем S в “длинном” окне

$$GM(m, f_k) = \sqrt[R]{\prod_{n=m-R+1}^m S(n, f_k)}$$

3. Среднее арифметическое по всем S в “длинном” окне

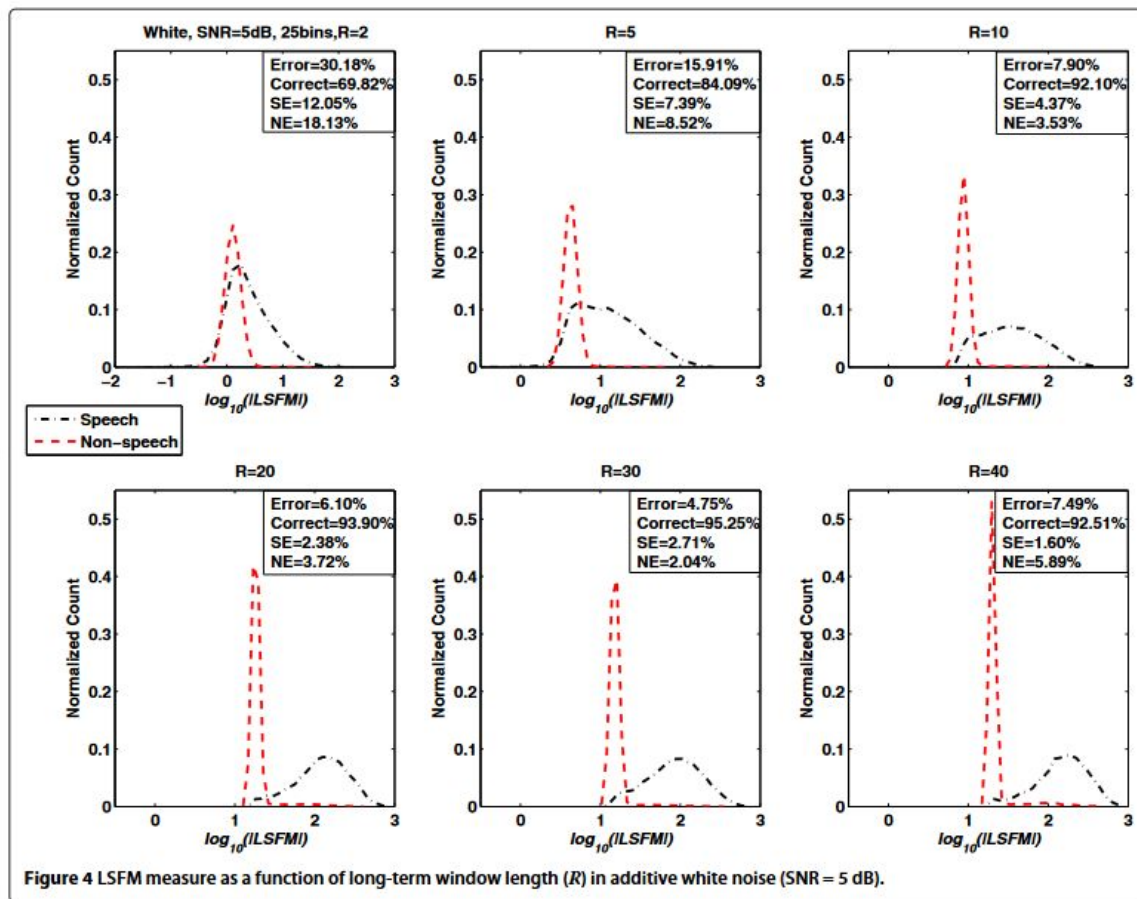
$$AM(m, f_k) = \frac{1}{R} \sum_{n=M-R+1}^m S(n, f_k)$$



Voice Activity Detection (Praat)

4. Вычисление LSFM

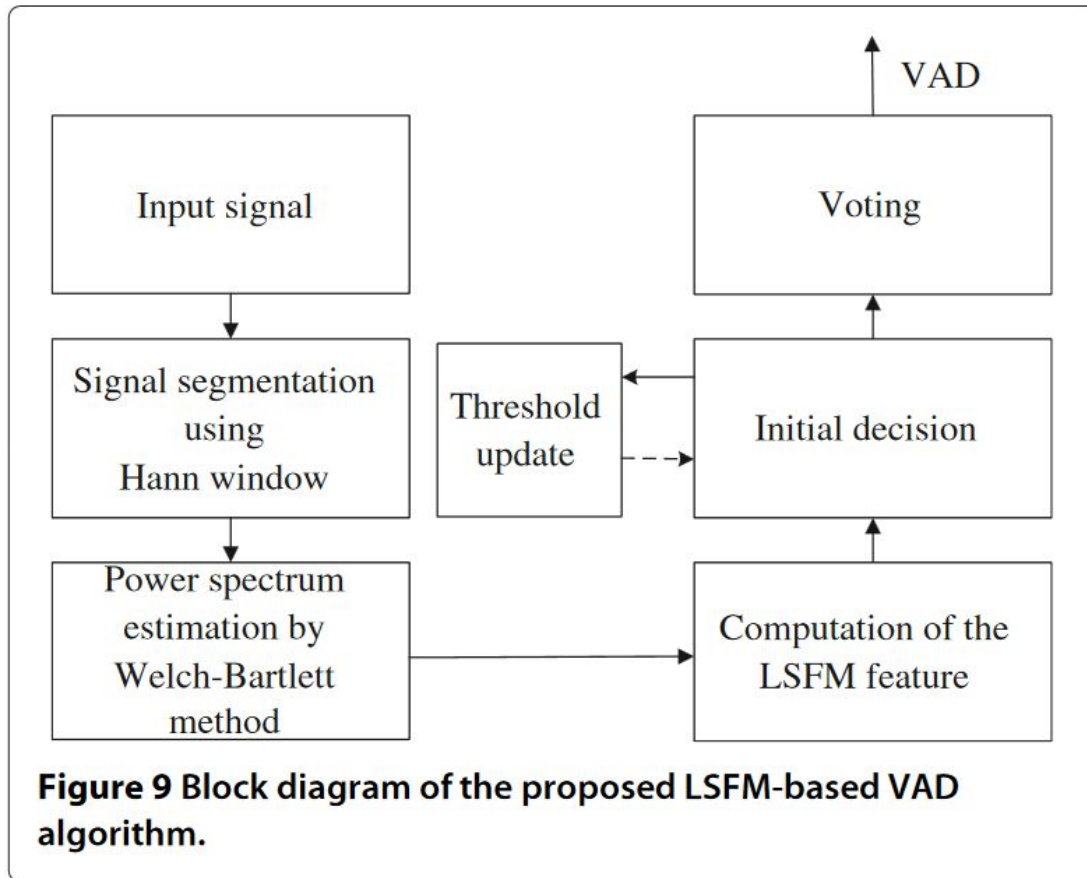
$$L_x(m) = \sum_k \log_{10} \frac{GM(m, f_k)}{AM(m, f_k)}$$





Voice Activity Detection (Praat)

5. Собственно алгоритм





Voice Activity Detection (Praat)

5.1. Первая часть сигнала (1.39 с для $R = 30$ и $M = 10$) считается неречевой.

5.2. Значение порога вычисляется как $\min(L_x)$ на этом промежутке

5.3. На каждом окне m ($N_w = 20$ мс, $N_{sh} = 10$ мс) обновляем порог следующим образом:

- 1) берём $\min(L_{speech})$ за последние 100 “длинных” речевых окон
- 2) берём $\max(L_{non-speech})$ за последние 100 “длинных” неречевых окон
- 3) складываем их с коэффициентами λ и $1-\lambda$ соответственно ($\lambda = 0.55$)

5.4. Для каждого окна получаем предварительные решения $V_{INL}(m)$: если в предыдущих R окнах есть хотя бы одно речевое окно, $V_{INL}(m) = 1$, иначе 0



Voice Activity Detection (Praat)

5.5. Делим сигнал на целевые промежутки = Nsh

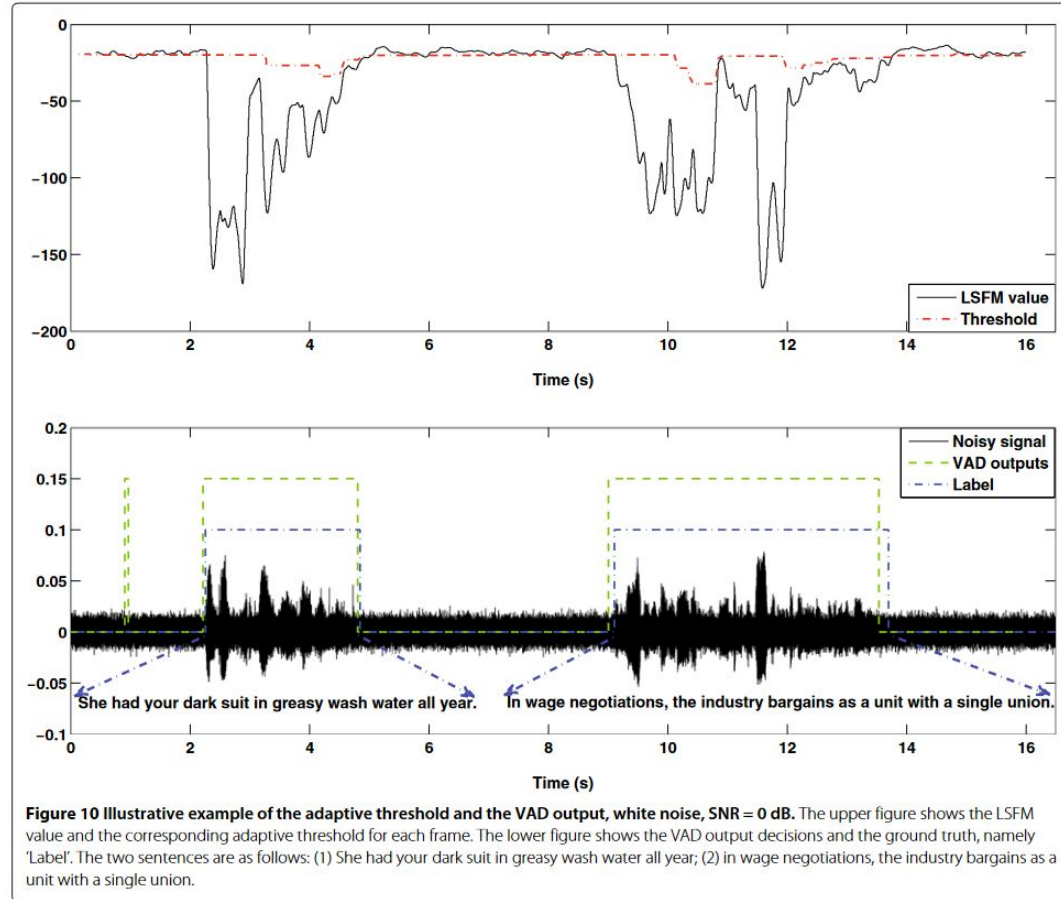
5.6. Для каждого промежутка определяем, какие окна пересекаются с ним, и для каждого из них собираем $V_{INL}(m)$, $V_{INL}(m + 1)$, ..., $V_{INL}(m + R - 1)$

5.7. Если среди собранных значений 80% речевые, отмечаем промежуток как речь

5.8. Иначе отмечаем его как паузу



Voice Activity Detection (Praat)





Оценка

1. Accuracy rate:

- а) CORRECT – доля правильно принятых решений
- б) speech hit rate (HR1) – доля правильно определённых речевых фрагментов
- в) non-speech hit rate (HR0) – доля правильно определённых неречевых фрагментов

2. Error rate:

- а) Front-end clipping (FEC) – начало речи, определённое как шум
- б) Mid-speech clipping (MSC) – фрагмент в середине речи, определённый как шум
- в) Noise detected as speech (NDS) – определение речи внутри паузы
- г) Carry over (OVER) – шум после конца речевого фрагмента, определённый как речь



Оценка

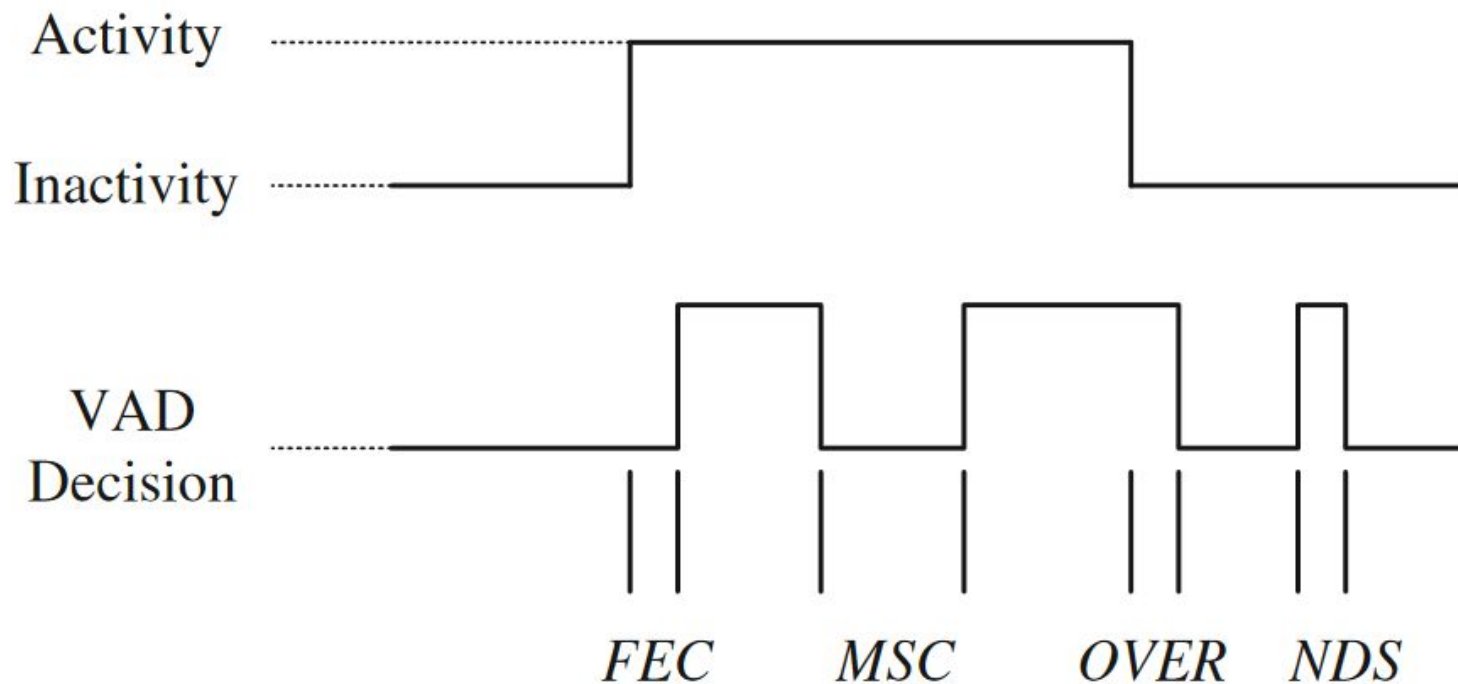


Figure 15 Objective parameters for performance evaluation.

Спасибо за внимание!

