Институт проблем искусственного интеллекта, г. Донецк

Классификация фреймов речевого сигнала в задачах дикторонезависимого распознавания речи

Ермоленко Т.В.

Определение границ речи в звуковом сигнале

1. Обучение шуму

Шум - сигнал $\varepsilon(n)$ (первые три поступивших на вход системы буфера данных)

$$\alpha(j,n)=M\left(E(j)\right)+n\sqrt{D\left(E(j)\right)} \qquad (1)$$

$E_k(j)$ – энергия вейвлет-спектра сигнала;

$M(E(j))$ и $D(E(j))$ – оценки математического ожидания и дисперсии энергии вейвлет-спектра шума $\varepsilon(n)$ на уровне j.

2. Определение границ речи

$$BOOL(k,n)=\begin{Bmatrix}1, & \left(\exists j_s\in M_s: E_k(j_s)>\alpha(j_s,n)\right)\vee\left(\exists j_v\in M_v: E_k(j_v)>\alpha(j_v,n)\right)\\ 0, & иначе\end{Bmatrix}$$

$E_k(j)$ – энергия вейвлет-спектра k-го фрейма сигнала;

$M_s=\left\{1,\ldots,j_s\right\}$ – множество масштабов, соответствует части спектра, в которой сосредоточена энергия шумных глухих щелевых или смычно-щелевых звуков;

$M_v=\left\{j_v,\ldots,j_{max}\right\}$ – множество масштабов, соответствует низкочастотной части спектра, в которой сосредоточена энергия вокализованных звуков.

$$\exists r>l:\forall i: r<i<r+m\left(BOOL(r,3)=1\right)\wedge\left(BOOL(i,3)=0\right)\wedge\left(m>maxPsL\right)\Rightarrow R=r\,\Delta N$$

$$\exists l:\forall i: l<i<l+m\left(BOOL(l,3)=0\right)\wedge\left(BOOL(i,3)=1\right)\wedge\left(m>minPnL\right)\Rightarrow L=l\,\Delta N$$

L и R – номера отсчетов сигнала, являющихся левой и правой границами слова,
minPnL – число фреймов, соответствующее минимальной длине фонемы;
maxPsL – число фреймов, соответствующее максимальной длине шумного глухого смычного звука.

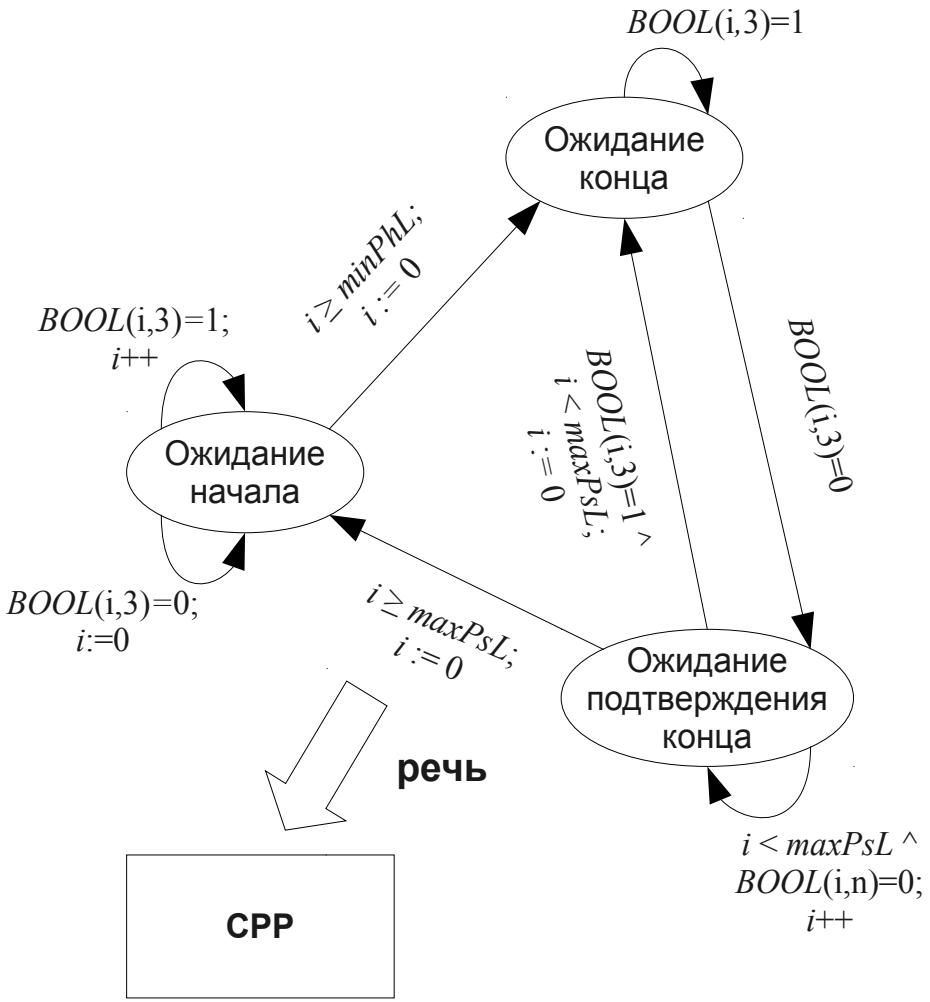

Рисунок 1 – Диаграмма состояний и переходов для определения границ речи

Классификация фреймов речевого сигнала

ШФК звуков речи: вокализованные (*Voc*); шумные глухие щелевые или смычно-щелевым (*Sh*); шумные глухие смычные (*P*).

$$P(k)=\frac{\sum_{j=HFBound}^{\Delta N-1}|FFT_j(k)|}{\sum_{j=0}^{HFBound-1}|FFT_j(k)|} \qquad (2),$$

HFBound – номер частоты, соответствующей левой границе высокочастотной части спектра, в которой сосредоточена энергия звуков из класса *Sh* (около 4 кГц);

$\{FFT_j\}_{j=0}^{\Delta N-1}$ – массив коэффициентов Фурье-спектра *k*-го фрейма.

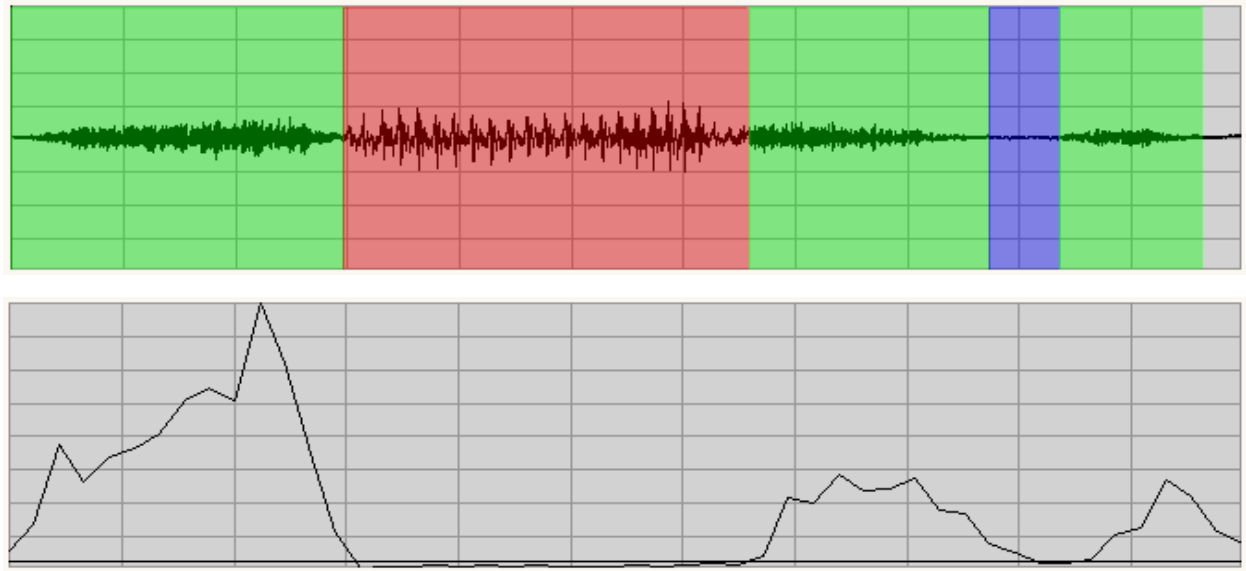

Рисунок 2 – Графики АВП (вверху) и значений $P(k)$ (внизу), полученные для реализации слова «шесть», горизонтальная линия соответствует значению порога $\alpha(n)=M(P)+n\sqrt{D(P)}$ при n=5

Для окончательной классификации на классы *Voc*, *Sh* и *P* используются значения вейвлет-спектра на множестве уровней разложения M_v . Значения энергий звуков классов *Sh* и *P* составляет менее 10% от энергии высокоамплитудных гласных.

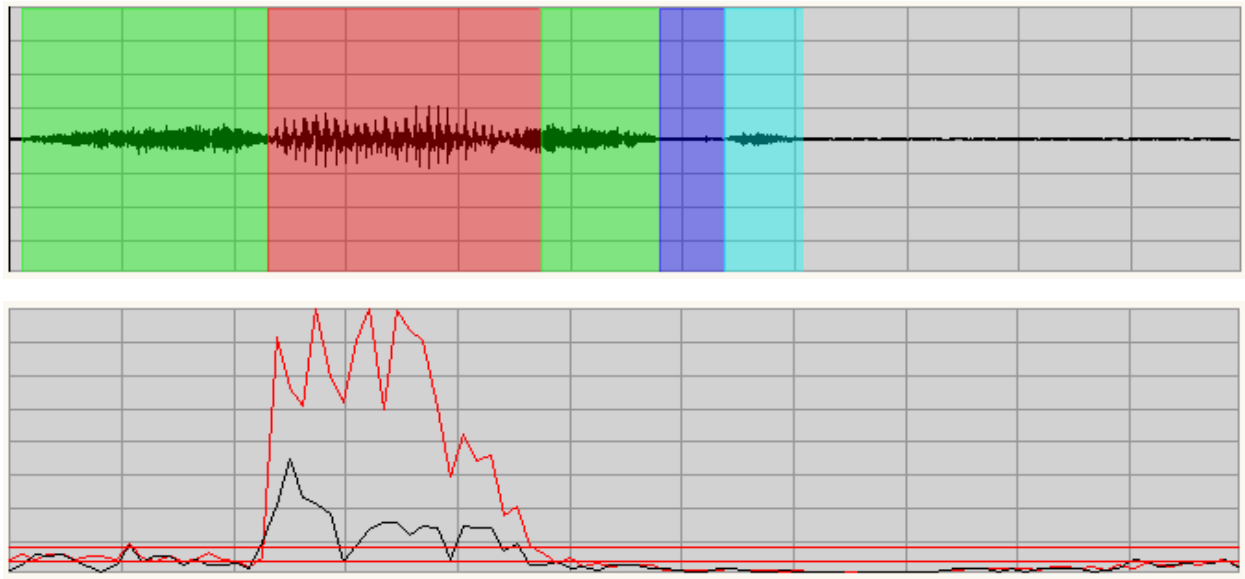

Рисунок 3 – Графики АВП (вверху) и коэффициентов вейвлет-спектра при *j*=5,6 (внизу), полученные для реализации слова «шесть», горизонтальные прямые соответствуют значению, составляющему 10% от максимального значения коэффициента вейвлет-спектра PC на соответствующих уровнях

Для классификации фреймов PC используется характеристика

$$BoolW(k)=\begin{Bmatrix}1,\ если\ \exists j_v\in M_v:\ E_k(j_v)>0.1\,maxLevEn(j_v)\\0,\ иначе\end{Bmatrix}$$

maxLevEn(*j*) – максимальное значение энергии коэффициента вейвлет-спектра на уровне *j*

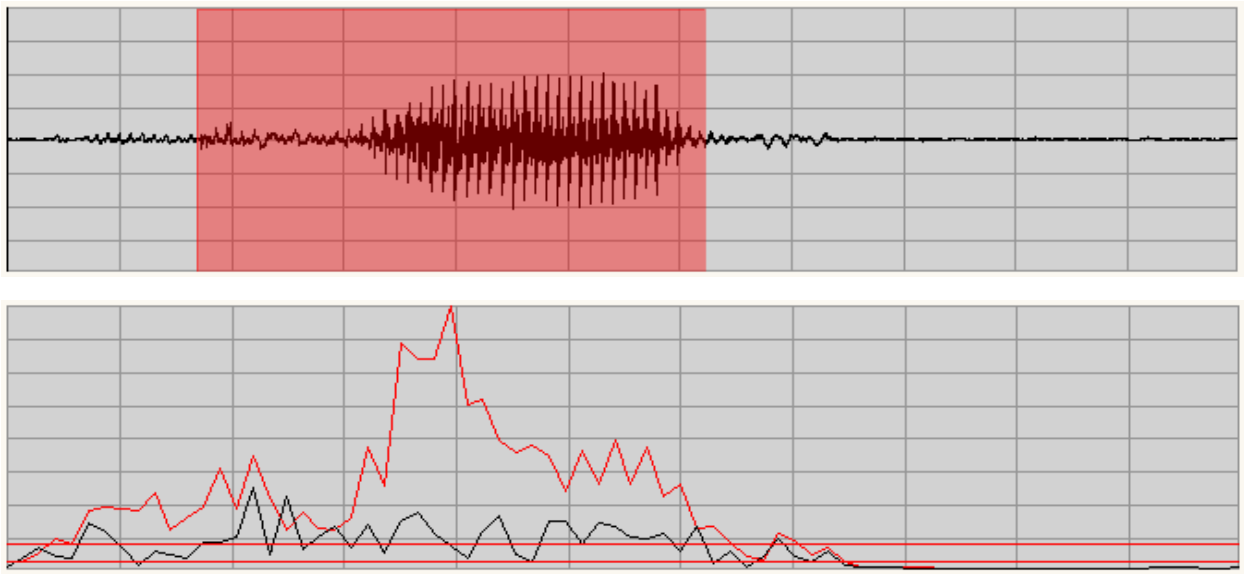

Рисунок 4 – Графики АВП (вверху) и коэффициентов вейвлет-спектра при *j*=5,6 (внизу), полученные для реализации слова «два», горизонтальные прямые соответствуют значению, составляющему 10% от максимального значения коэффициента вейвлет-спектра PC на соответствующих уровнях

Набор решающих правил для классификации фреймов:

$$P(k) \geq \alpha(n) \Rightarrow k \in Sh \vee k \in Voc \quad , \quad P(k) < \alpha(n) \Rightarrow k \in P \vee k \in Voc$$

$$BoolW(k) = 0 \Rightarrow k \in Sh \vee k \in P \quad , \quad BoolW(k) = 1 \Rightarrow k \in Voc$$

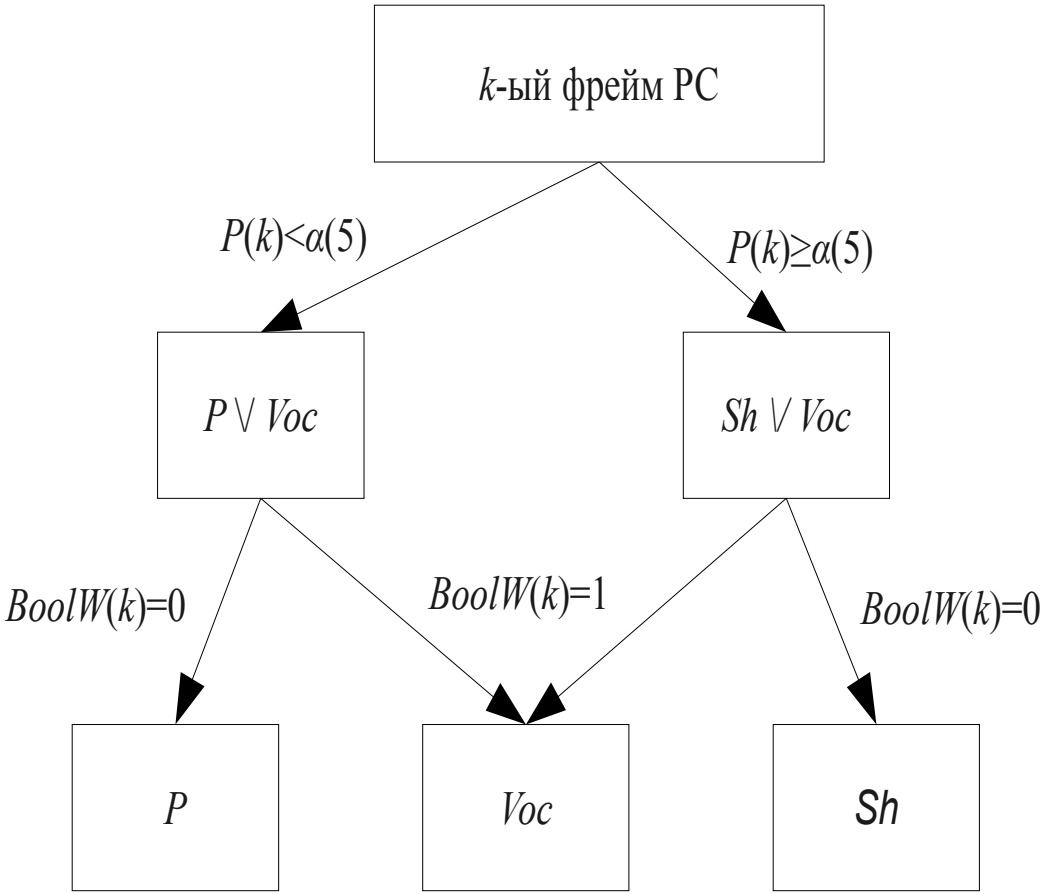

Рисунок 5 – Дерево решений для классификации фреймов РС

Спасибо за внимание :)