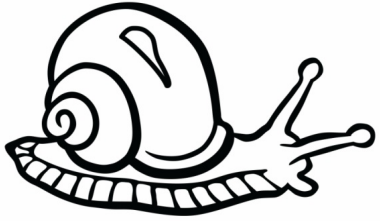
Автоматическая сегментация речевого сигнала

П. А. Холявин

p.kholyavin@spbu.ru

28.11.2024



Методы сегментации

1. С опорой на транскрипцию (forced alignment)

2. Без опоры на транскрипцию

1. По правилам

2. Статистические

Автоматическое определение взрывных

Признаки:
1. Общая энергия
2. Энергия выше 3 кГц
3. Spectral flatness

С шагом 1 мс, окно 5 мс

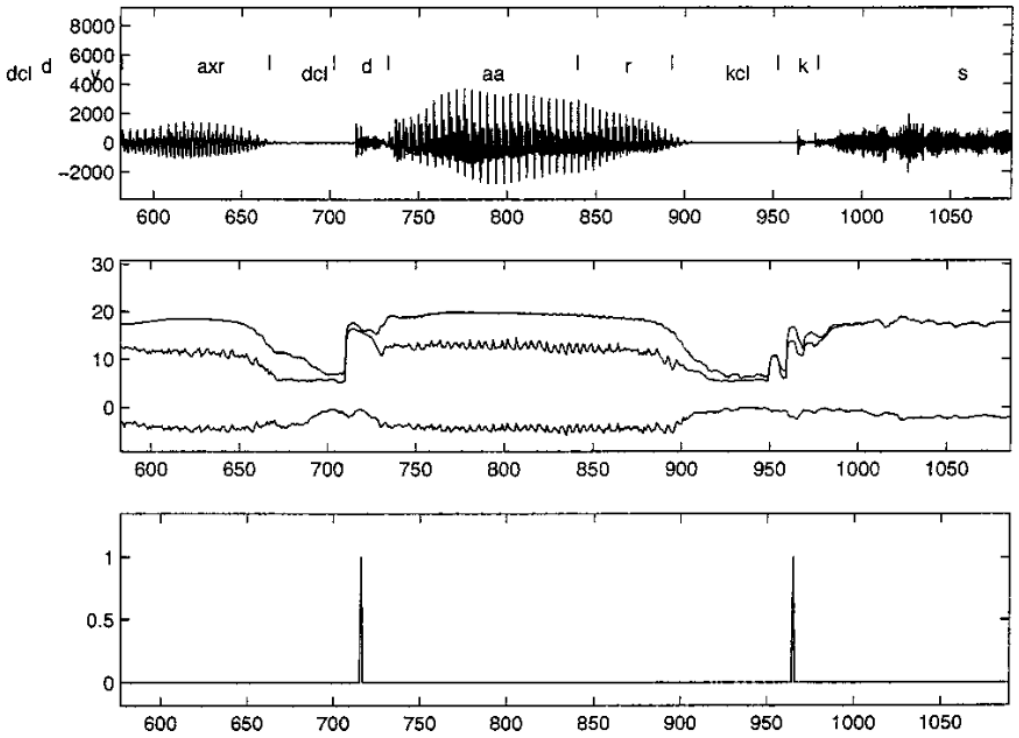


FIG. 1. Portion of the speech wave form $s(n)$ (top panel), the associated three-dimensional feature vector, $\mathbf{x}(n)$ (middle panel), and the desired output $y(n)$ bottom panel marking the times of the closure–burst transition.

Автоматическое определение взрывных

Алгоритмы:
A. Сумма разностей первого и второго признака

B. Оптимальный (обученный) оператор с двумя признаками

C. -//- с тремя признаками

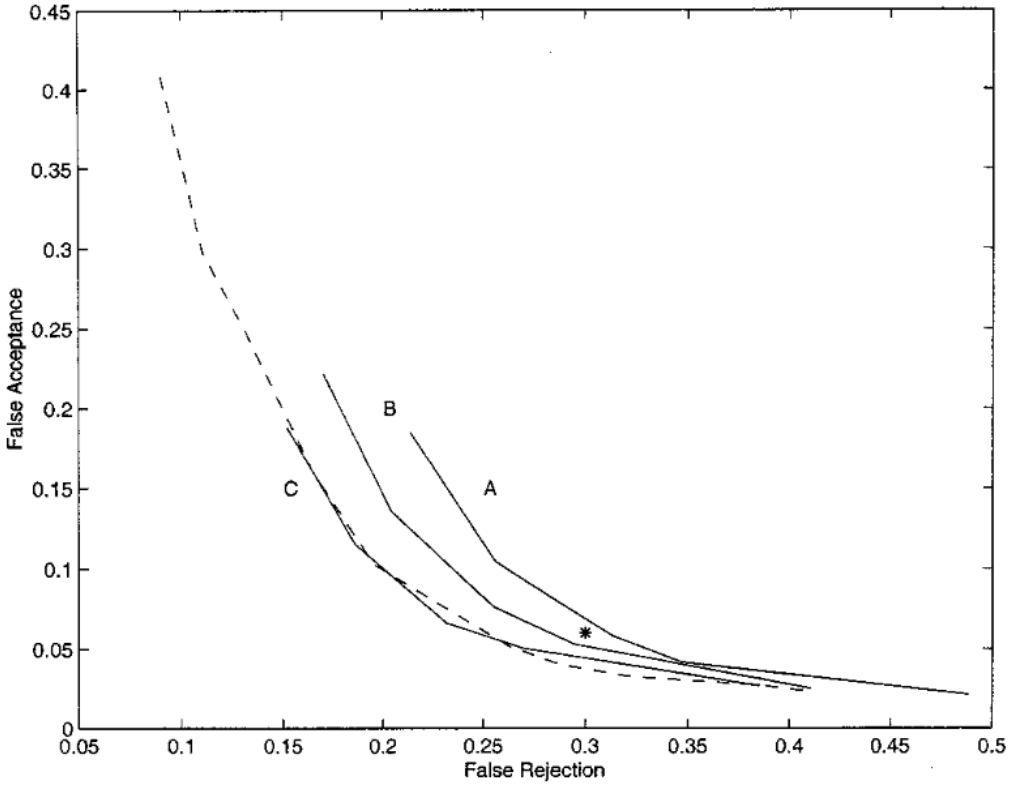
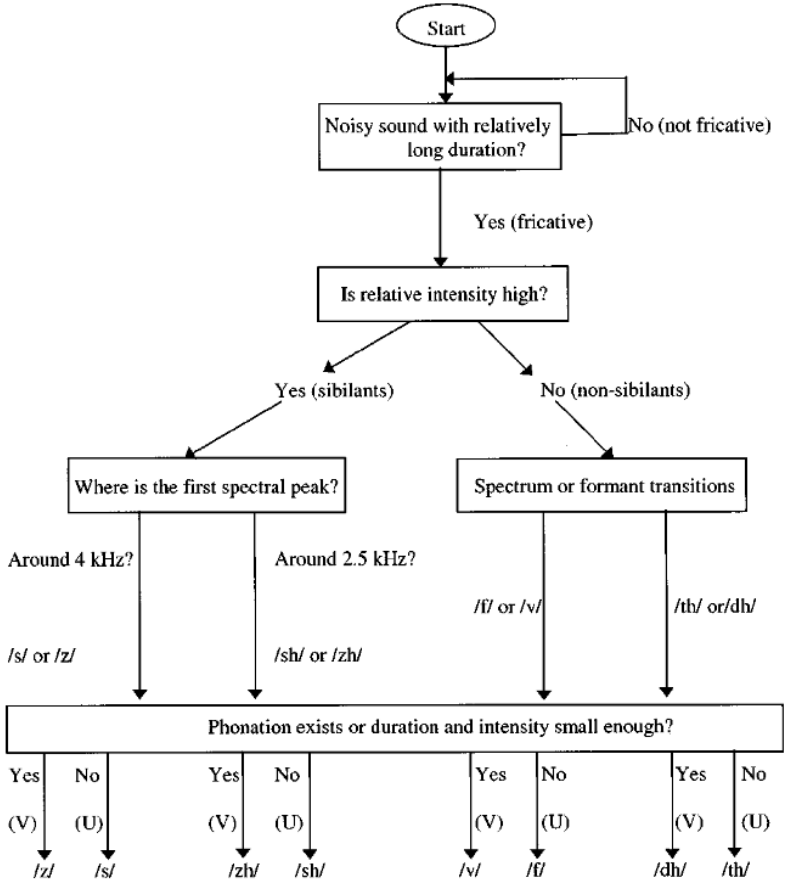


FIG. 2. ROC curves for detection of stop consonants using three different algorithms described in text.

Классификация фрикативных



Классификация фрикативных

Признаки для определения звонкости:

LOWG – энергия до 1 кГц

LOWE – отношение энергий частот до 1.5 кГц и от 3 кГц

DUP – длительность глухого участка

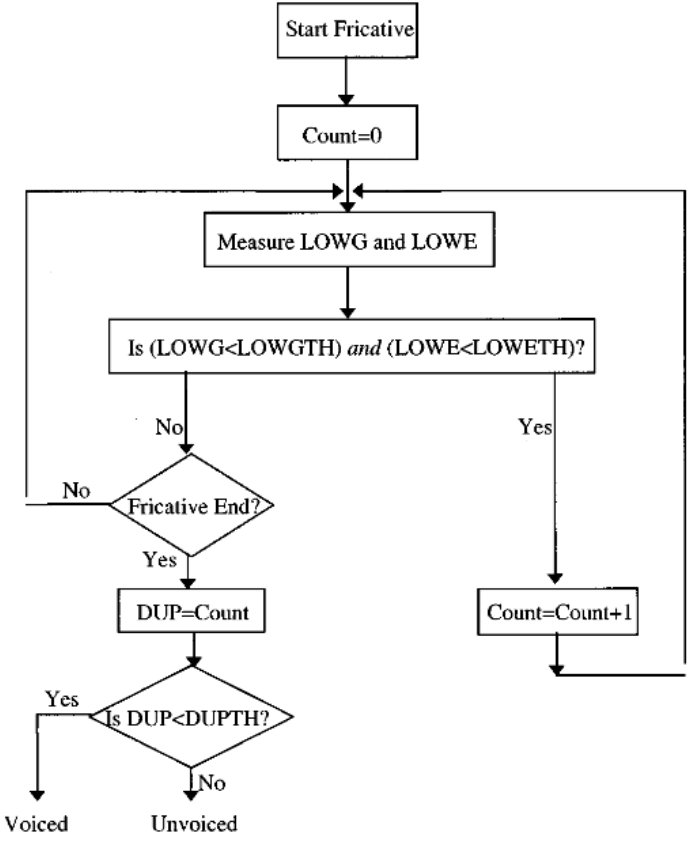


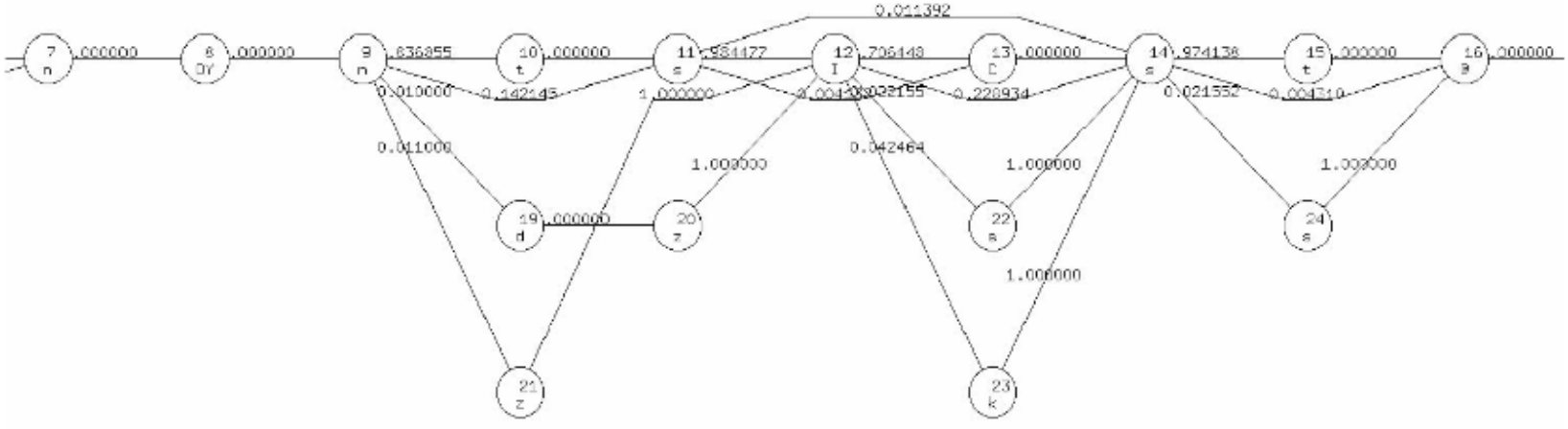
FIG. 5. Voicing detection in prevocalic fricatives.

Munich Automatic Segmentation System

С опорой на орфографию:

1. G2P

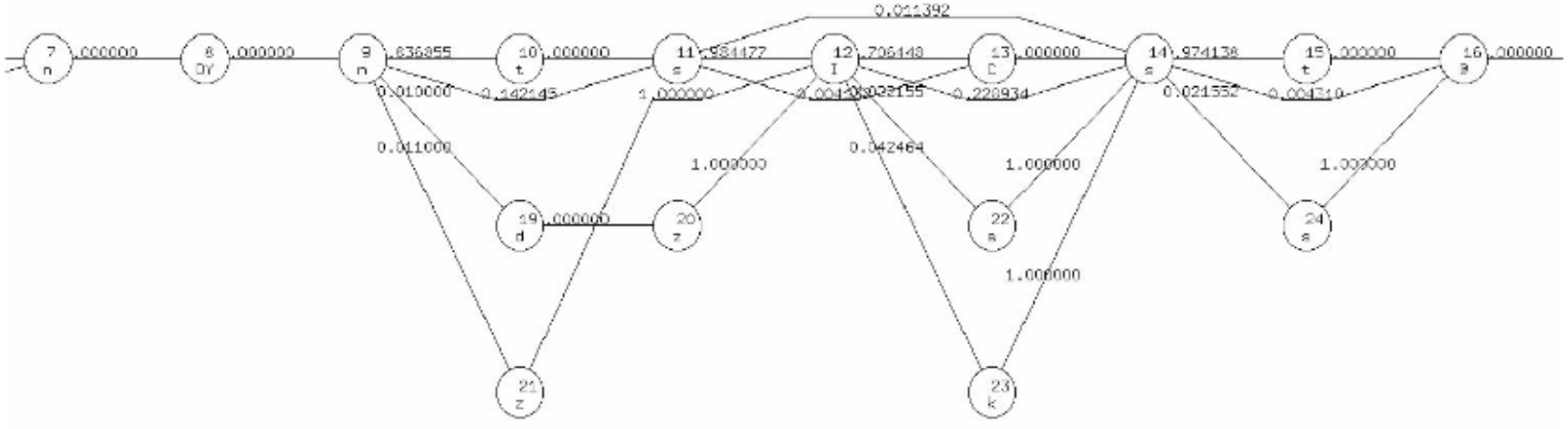
2. WORDVAR – построение графа произносительных вариантов



Munich Automatic Segmentation System

С опорой на орфографию:

3. Алгоритм Витерби (поиск наиболее вероятного пути через граф)



Montreal Forced Aligner

Kaldi

Признаки:

13 MFCC до 8 кГц + дельта и дельта-дельта, окно 25 мс, шаг 10 мс

Обучение:

1. Монофонные GMM-модели (40 итераций, 20 из них с пересчётом границ)
2. Трифонные GMM-модели (35 итераций, 15 с пересчётом)

Уточнение границ после HMM

1. На основе спектральных признаков

2. С помощью SVM

3. С помощью нейронных сетей

...

Моделирование границ

Использование фонетической сегментации:

Table 2. *Agreement percentages for different tolerances (in ms), for systems using or not using manual segmentation for training monophone HMMs.*

	<10	<20	<30	<40	<50
Segmentation not used for training	70.20	89.98	95.74	97.88	98.92
Segmentation used for training	**73.23**	**91.85**	**96.45**	**98.17**	**99.05**

Моделирование границ

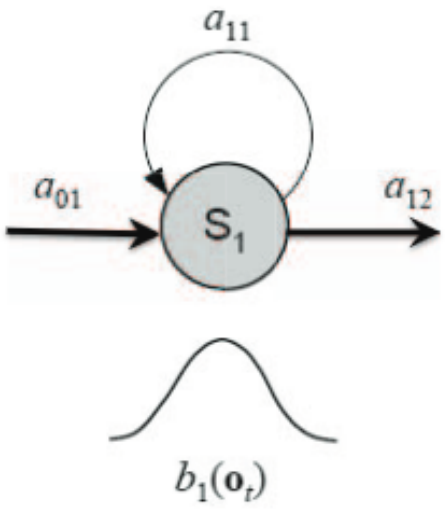


Figure 1: *1-state HMM. The one state HMM is a special 1-state model for the boundaries when the transition probabilities $a_{11} = 0$ and $a_{12} = 1$.*

Моделирование границ

Table 4. *Agreement percentages for different tolerances (in ms), for systems using monophone HMMs, monophone HMMs and boundary models, triphone HMMs, and triphone HMMs and boundary models.*

	<10	<20	<30	<40	<50
Monophones	73.23	91.85	96.45	98.17	99.05
Monophones & Bo.	77.44	**93.92**	**97.43**	**98.78**	99.35
Triphones	74.93	92.37	96.72	98.33	99.09
Triphones & Bo.	**78.09**	93.85	97.37	98.72	**99.37**

Использование Wav2Vec2

Table 1. Evaluation results of text-dependent alignment

Model	P	R	F1	R-val	Overlap
FAVE	0.57	0.59	0.58	0.64	74.3%
MFA-Libris	0.61	0.61	0.61	0.67	73.5%
MFA	0.62	0.63	0.63	0.68	75.0%
Gentle	0.49	0.46	0.48	0.56	67.7%
WebMAUS	**0.70**	**0.70**	**0.70**	**0.75**	78.8%
W2V2-FC-20ms-Libris	0.49	0.47	0.48	0.56	73.8%
W2V2-FC-10ms-Libris	0.57	0.54	0.55	0.62	76.4%
W2V2-FC-32k-Libris	0.66	0.63	0.64	0.69	79.3%
W2V2-FS-20ms	0.47	0.49	0.48	0.55	71.6%
W2V2-FS-10ms	0.68	0.68	0.68	0.73	**80.4%**
W2V2-FS-32k	0.63	0.65	0.64	0.69	79.3%
Pretrained G2P converter					
W2V2-FS-20ms	0.40	0.42	0.41	0.49	65.1%
W2V2-FS-10ms	0.56	**0.58**	0.57	0.63	72.5
W2V2-FC-32k-Libris	**0.58**	0.57	0.58	**0.64**	**73.0%**
Phone set adaptation (TIMIT-61)					
W2V2-FS-20ms	0.49	0.53	0.51	0.57	70.5%
W2V2-FS-10ms	**0.66**	**0.70**	**0.68**	**0.72**	**79.7%**

Использование Wav2Vec2

Table 2. Evaluation results of text-independent alignment

Model	P	R	F1	R-val	Overlap
W2V2-CTC-10ms	0.31	0.29	0.30	0.42	43.9%
W2V2-CTC-20ms	0.31	0.30	0.31	0.42	46.6%
Phone recognition + W2V2-FS					
W2V2-FS-20ms	0.40	0.42	0.41	0.48	64.2%
W2V2-FS-10ms	0.56	0.58	0.57	0.63	71.5%
W2V2-FC-32k-Libris	0.57	0.57	0.57	0.64	72.2%
Direct inference					
W2V2-FC-20ms-Libris	0.57	0.59	0.58	0.63	72.7%
W2V2-FC-10ms-Libris	0.55	0.58	0.56	0.62	72.5%
W2V2-FC-32k-Libris	**0.60**	**0.63**	**0.61**	**0.66**	**74.3%**

Спасибо за внимание!