Speech Technology Center

COMBINING HMM AND UNIT SELECTION TECHNOLOGIES TO INCREASE NATURALNESS OF SYNTHESIZED SPEECHDATA-DRIVEN

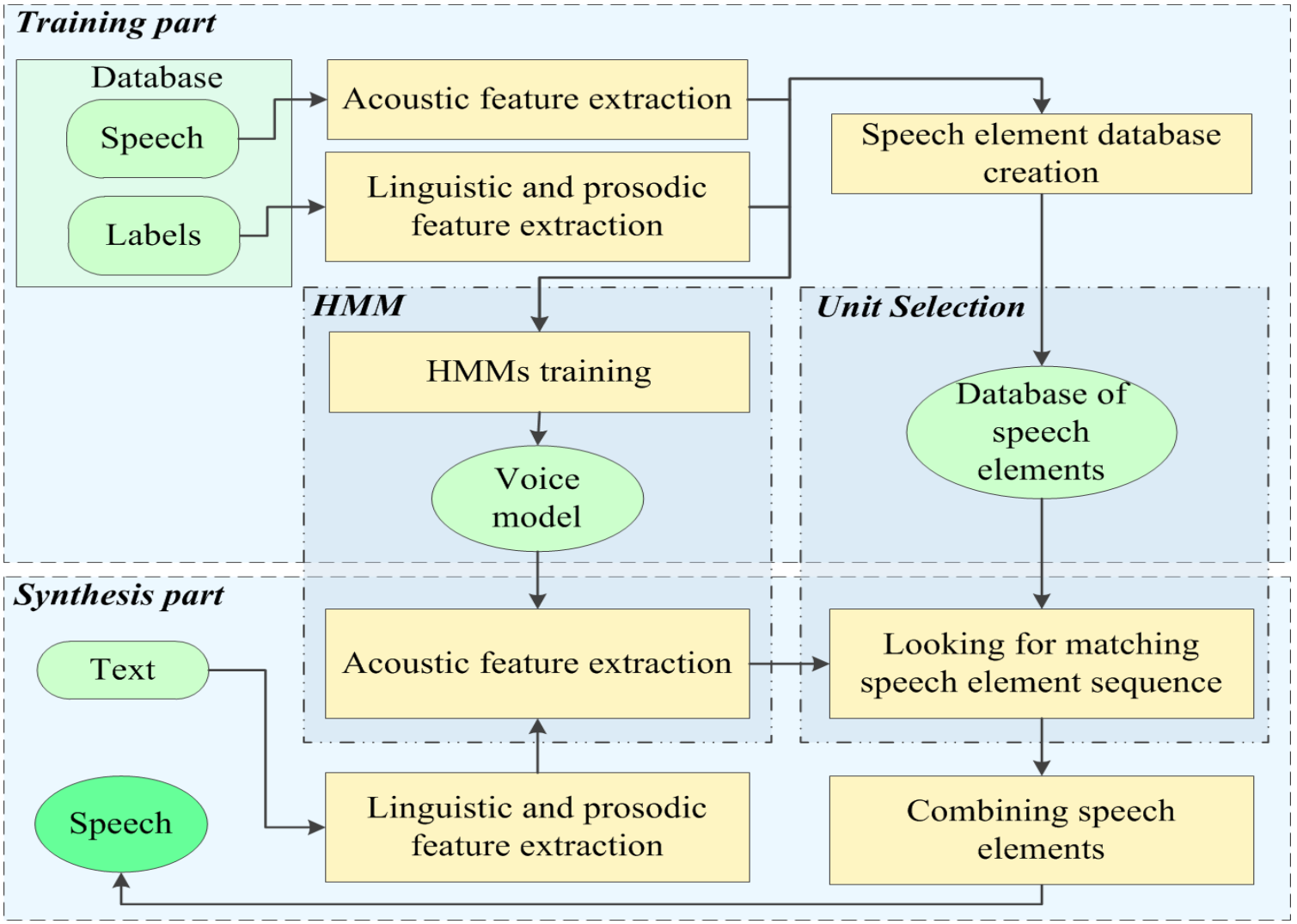
Chistikov Pavel
Korolkov Evgeny, Talanov Andrey
{chistikov,korolkov,andre}@speechpro.com
01.06.2013

Objectives

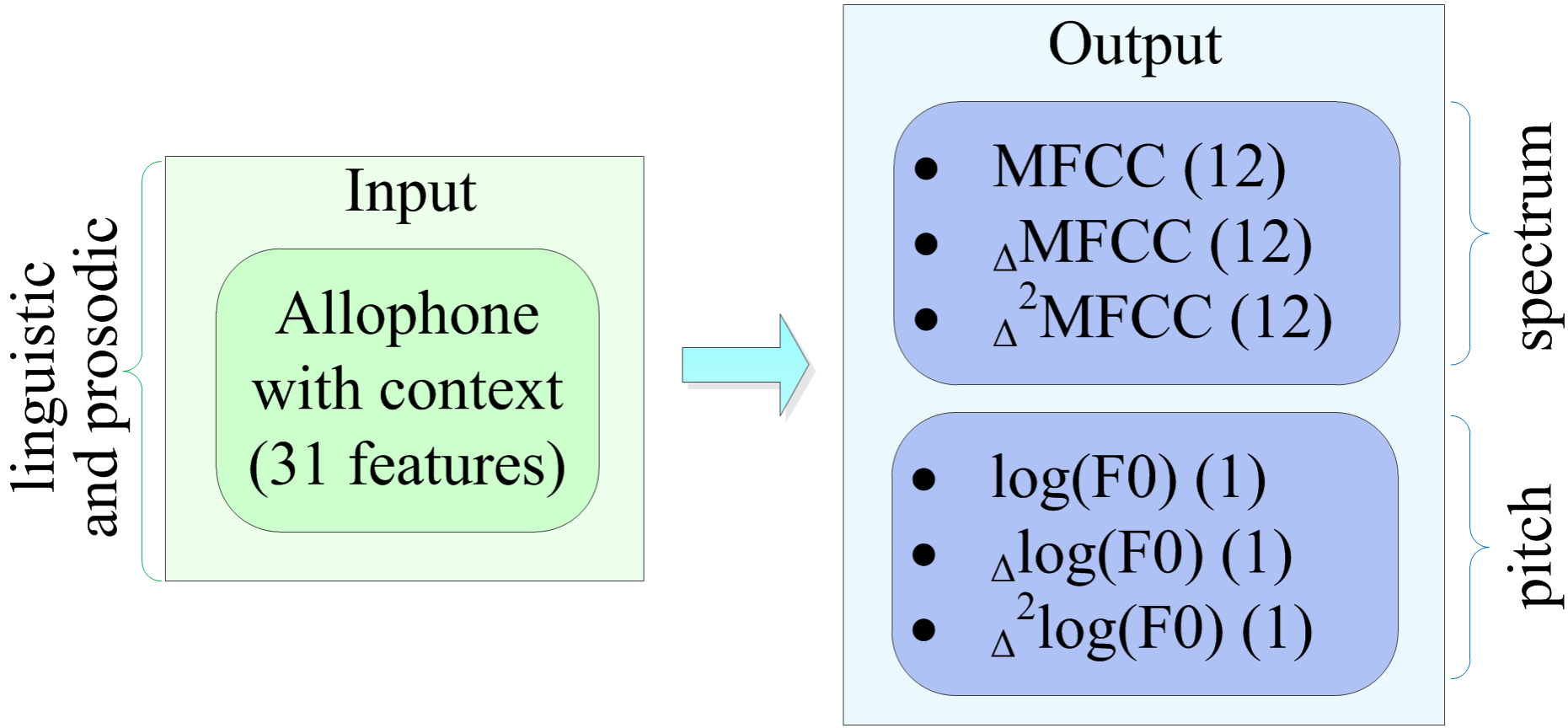
We propose a text-to-speech system based on the two most popular approaches: statistical speech synthesis (based on hidden Markov models) and concatenative speech synthesis (based on Unit Selection). This approach:

- improves the quality of synthesized speech;
- speeds up the process of new voice creation.

The basic steps conducted by the TTS engine



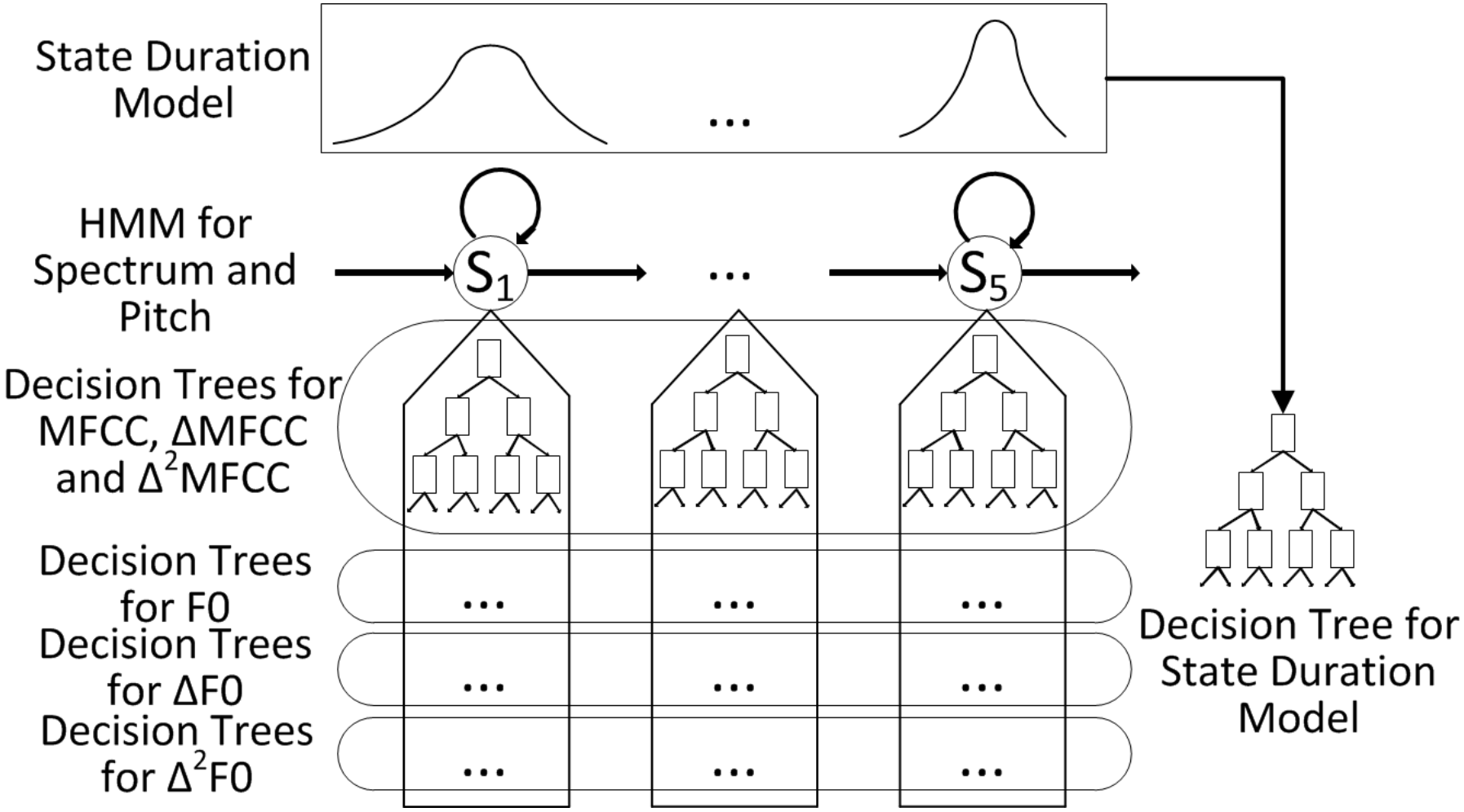
Modeled features



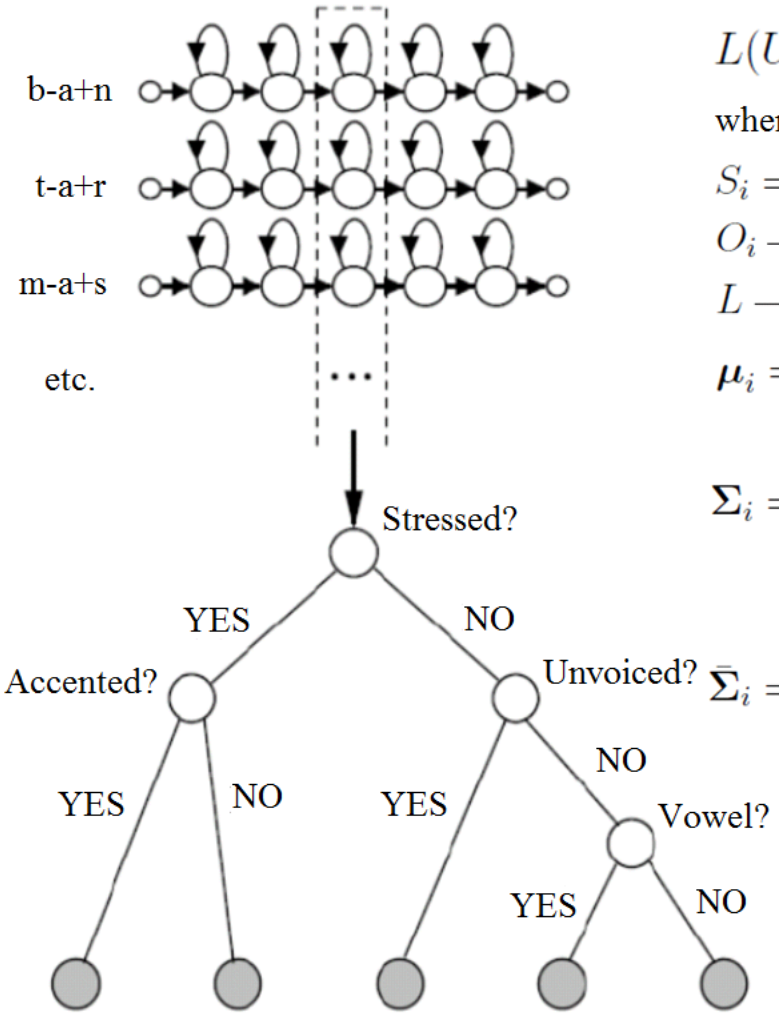
Contextual features

Allophone features - 7
Phone before previous
Previous phone
Current phone
Next phone
Phone after next
Phone position from the beginning of the syllable
Phone position from the end of the syllable
Syllable features - 13
Word features - 8
Sentence features - 3

Voice model


State Duration Model
HMM for Spectrum and Pitch
Decision Trees for MFCC, ΔMFCC and Δ²MFCC
Decision Trees for F0
Decision Trees for ΔF0
Decision Trees for Δ²F0
S1
S5
Decision Tree for State Duration Model

State clustering



$$L(U) = -\frac{1}{2}T\left(L + L\log 2\pi + \log|\bar{\Sigma}|\right) \to \max,$$

where U — set of states $\{S_1, S_2, ..., S_M\}$,

$S_i = N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $T = \sum_{i=1}^{M} O_i$,

O_i — number of state repetitions S_i,

L — feature vector size,

$\boldsymbol{\mu}_i = \{\mu_{i1}, ..., \mu_{iL}\}$,

$$\boldsymbol{\Sigma}_i = \begin{Bmatrix} \sigma_{i1}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{iL}^2 \end{Bmatrix}, \quad \bar{\sigma}_i^2 = (\hat{\sigma}_i^2 - \hat{\mu}_i/T)/T,$$

$$\hat{\mu}_i = \sum_{j=1}^{M} \mu_{ij} O_j,$$

$$\bar{\boldsymbol{\Sigma}}_i = \begin{Bmatrix} \bar{\sigma}_{i1}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \bar{\sigma}_{iL}^2 \end{Bmatrix}, \quad \hat{\sigma}_i^2 = \sum_{j=1}^{M} (\sigma_{ij}^2 + \mu_{ij}^2) O_j.$$

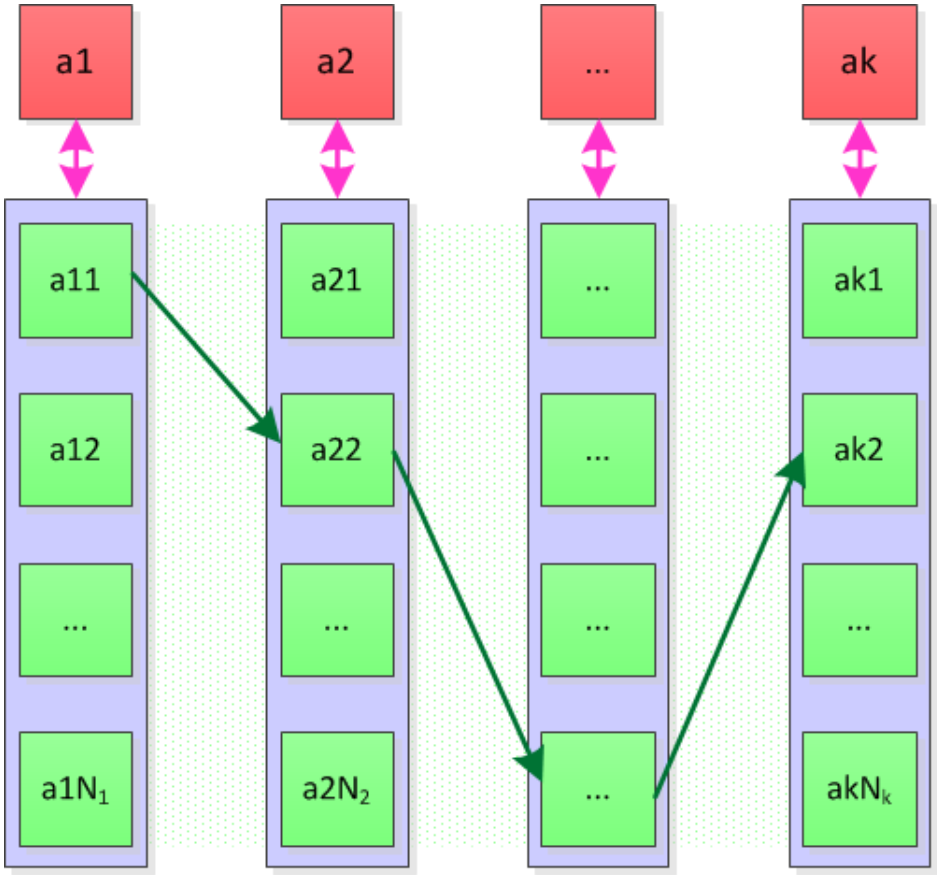
Unit Selection method

The task is to select a set of allophones $u_1,u_2,\ldots,u_n$ from the database which minimizes the cost function (1).

$$C(u,t) = \sum_{i=1}^{n} C^t(u_i,t_i) + \sum_{i=2}^{n} C^c(u_{i-1},u_i), \quad (1)$$

where n is the number of elements in the sequence; t is the number of allophones in the database; C^t is the target cost; C^c is the concatenation cost.

Speech synthesis scheme



Target cost criterions:
- Pitch
- Energy
- Duration
- Context

Concatenation cost criterions:
- Pitch
- Pitch delta
- MFCC
- Continuity

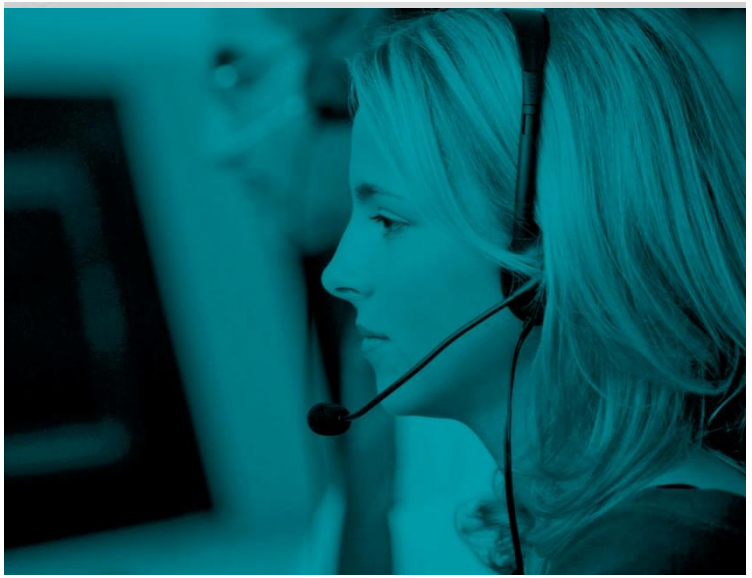
- target element

- element-candidate

Conclusions

- Speech parameters are obtained from HMMs whose observation vectors consist of spectrum, F0 and duration features.
- Context-clustering is performed to achieve a greater flexibility of the algorithm and to enable the use of the voice model even when a small database is used.
- Speech elements are chosen from the database based on parameters modeled by Unit Selection method.
- Experimental results show that combining HMM and Unit Selection methods improves the quality of synthesized speech.

Thank you for your attention



ABOUT THE COMPANY

ABOUT THE COMPANY

Speech Technology Center (STC) is an international leader in speech technology and multimodal biometrics. It has over 20 years of research, development and implementation experience in Russia and internationally.

STC is a leading global provider of innovative systems in high-quality recording, audio and video processing and analysis, speech synthesis and recognition, and real-time, high-accuracy voice and facial biometrics solutions. STC innovations are used in both public and commercial sectors, from small expert laboratories, to large, distributed contact centers, to nation-wide security systems.

STC is ISO-9001: 2008 certified.

CONTACTS

Russia
4 Krasutskogo street, St. Petersburg, 196084
Tel.: +7 812 331 0665
Fax: +7 812 327 9297
Email: info@speechpro.com

USA
Suite 316, 369 Lexington ave
New York, NY, 10017
Tel.: +1 646 237 7895
Email: sales-usa@speechpro.com