REAL-TIME VOICE CONVERSION USING ARTIFICIAL NEURAL NETWORKS WITH RECTIFIED LINEAR UNITS

E. Azarov, M. Vashkevich, D. Likhachov and A. Petrovsky

Department of Computer Engineering,

Belarusian State University of Informatics and Radioelectronics,

Minsk, Belarus

Work was supported by IT-MOBILE company (Moscow)

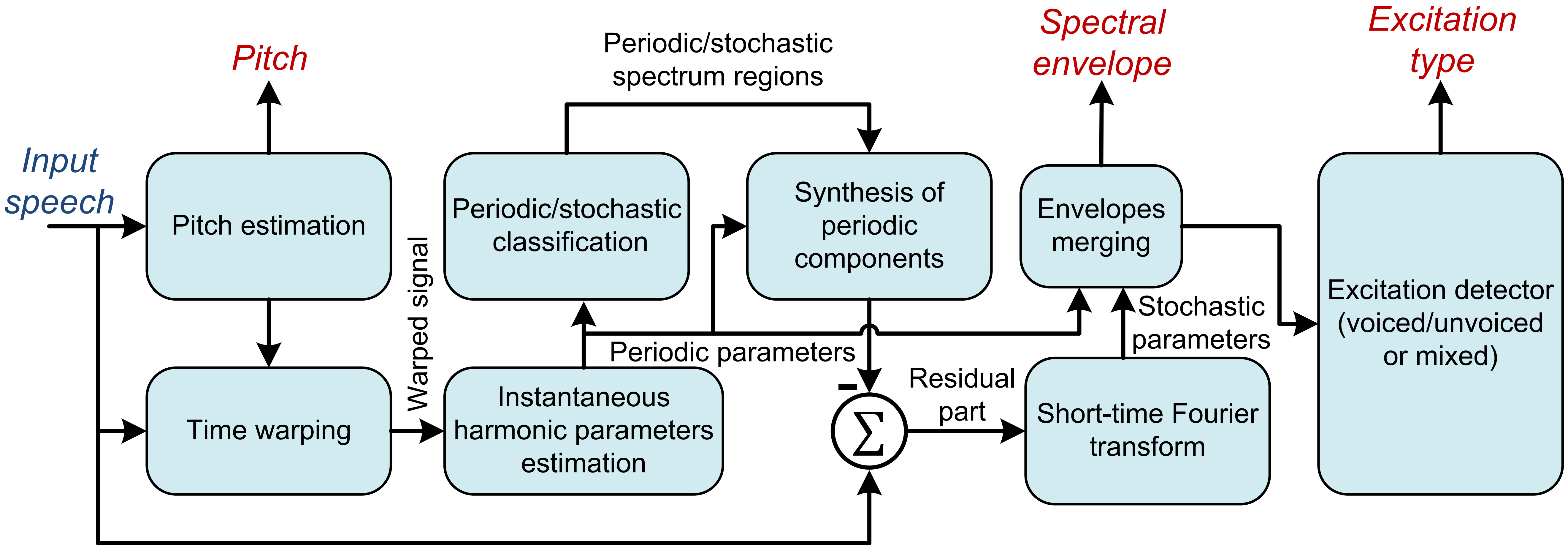
1. Introduction

This paper presents an approach to parametric voice conversion that can be used in real-time entertainment applications.

Main features:

- spectral mapping is implemented using an artificial neural network (ANN) with rectified linear units (ReLU);
- the oversmoothing effects is reduced by a special network configuration and utilization of temporal states of the speaker;
- the speech signal is represented using the harmonic plus noise model with multicomponent sinusoidal excitation;
- signal processing is performed in warped time domain;

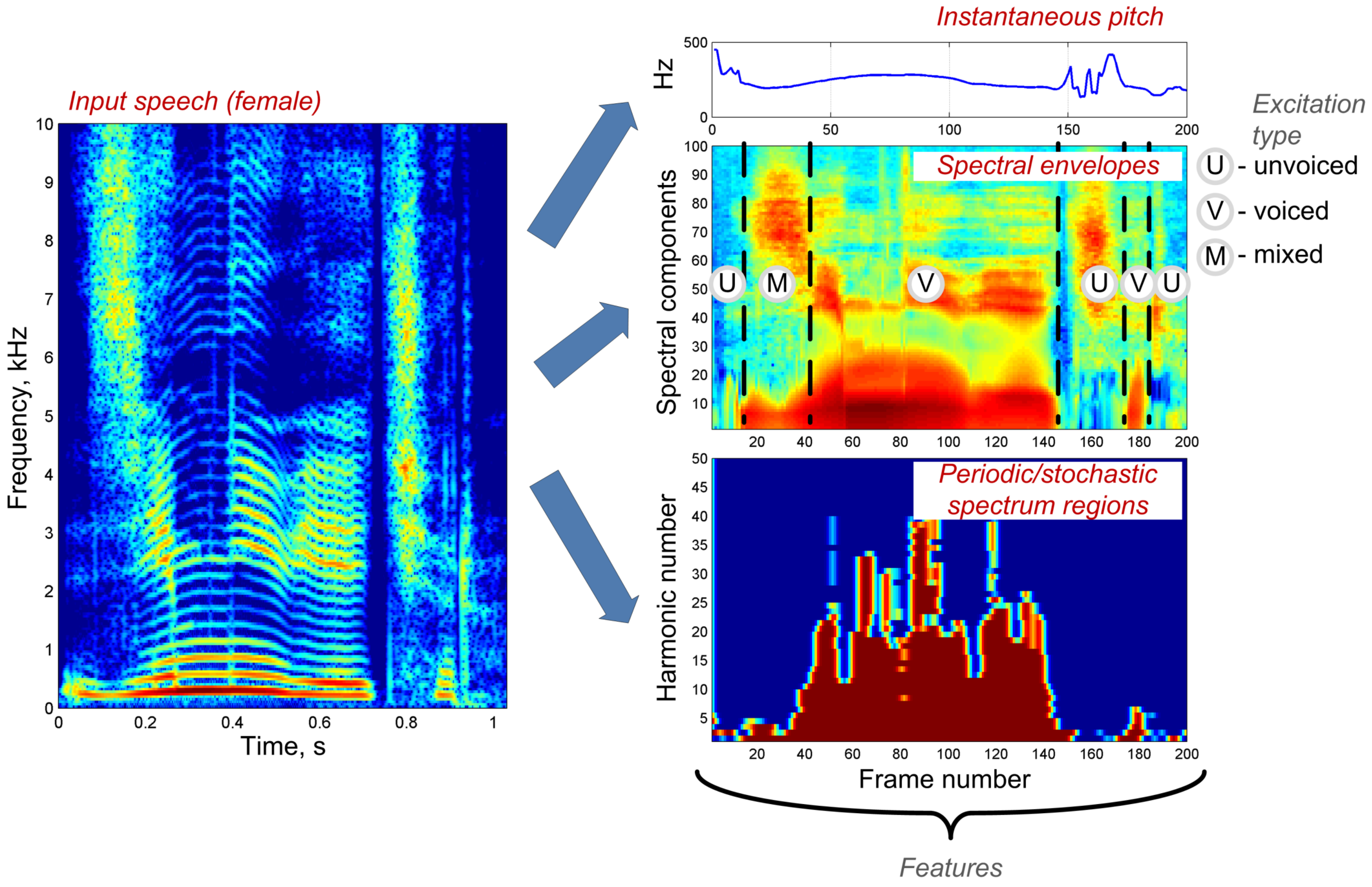
2. Feature extraction routine


Pitch
Periodic/stochastic spectrum regions
Spectral envelope
Excitation type
Input speech
Pitch estimation
Periodic/stochastic classification
Synthesis of periodic components
Envelopes merging
Excitation detector (voiced/unvoiced or mixed)
Warped signal
Stochastic parameters
Periodic parameters
Time warping
Instantaneous harmonic parameters estimation
-
Σ
Residual part
Short-time Fourier transform

Algorithm: 1) instantaneous pitch values are extracted using the instantaneous robust algorithm for pitch tracking[1] (IRAPT); 2) the signal is transformed into warped time domain to get the signal with constant pitch; 3) instantaneous harmonic parameters are estimated using a DFT-modulated filter bank; 4) spectrum regions are classified as periodic or stochastic; 5) periodic components are synthesized and subtracted from the signal; 6) the residual is transformed into frequency domain using the STFT; 7) the estimated instantaneous harmonic parameters and the residual spectrum are combined into joint spectral envelope; 8) adjacent spectral envelopes are analyzed by excitation detector that makes the whole frame decision – voiced/unvoiced or mixed.

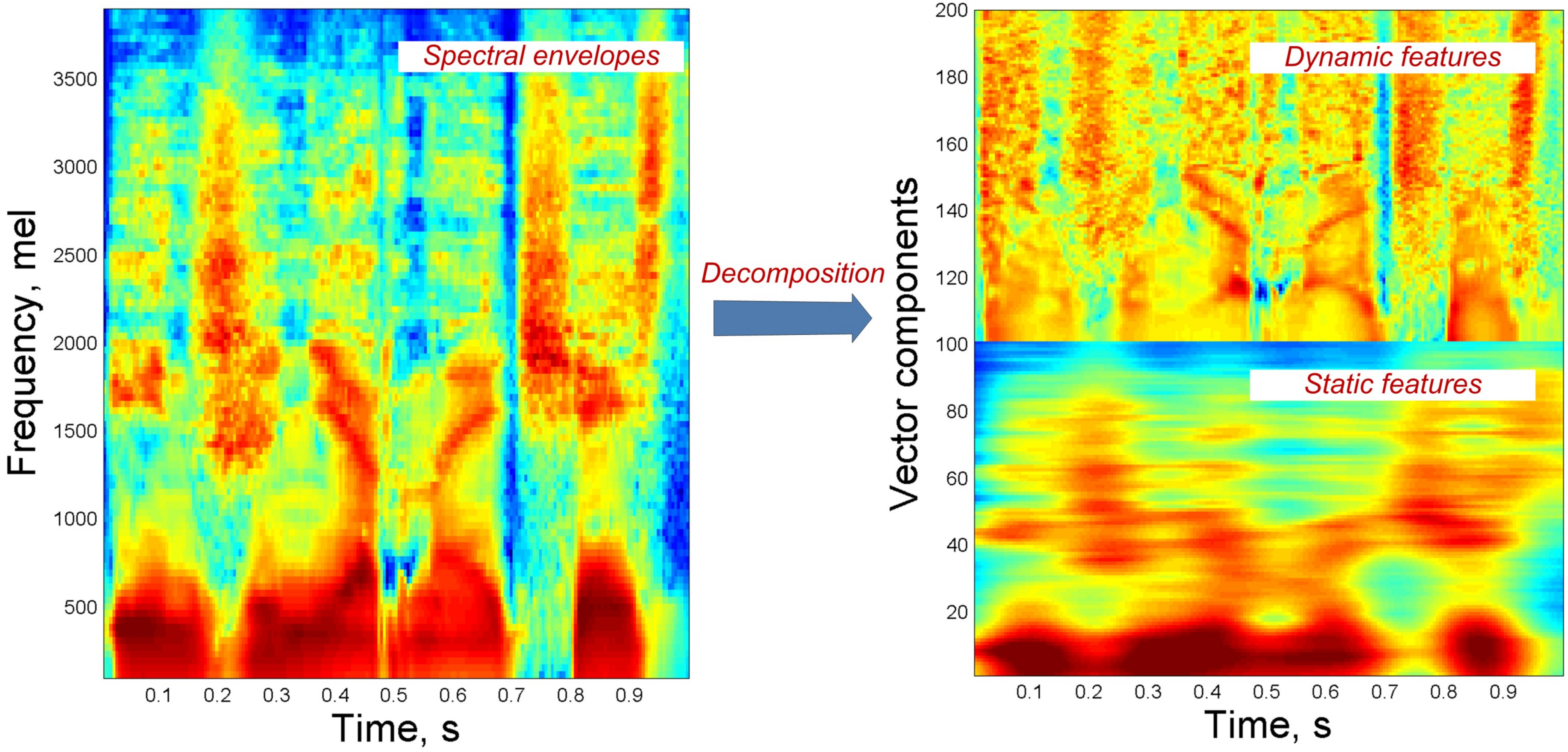
[1] Azarov, E., Vashkevich, M., and Petrovsky A., “Instantaneous pitch estimation based on RAPT framework,” *Proc. EUSIPCO*, Bucharest, Romania, Aug. 2012.

3. Feature extraction: example



Instantaneous pitch, spectral envelope, periodic/stochastic spectrum regions and excitation labels are extracted from the input female speech.

4. Static and dynamic features



Static and dynamic features are obtained by decomposing time series of envelopes into low-frequency and high-frequency parts using a low-pass filter with cut-off frequency 4-9Hz. The low frequency amplitude modulations capture speaker specific features when high frequency modulations contain phonetical and articulatory information.

5. Training the ANN (1)

Mapping function

$$Y = w_5 \mathrm{RL}\left(w_4 \begin{bmatrix} \mathrm{RL}(w_1 X + bias_1) \\ w_2 S_s \\ w_3 S_t \end{bmatrix} + bias_4 \right) + bias_5$$

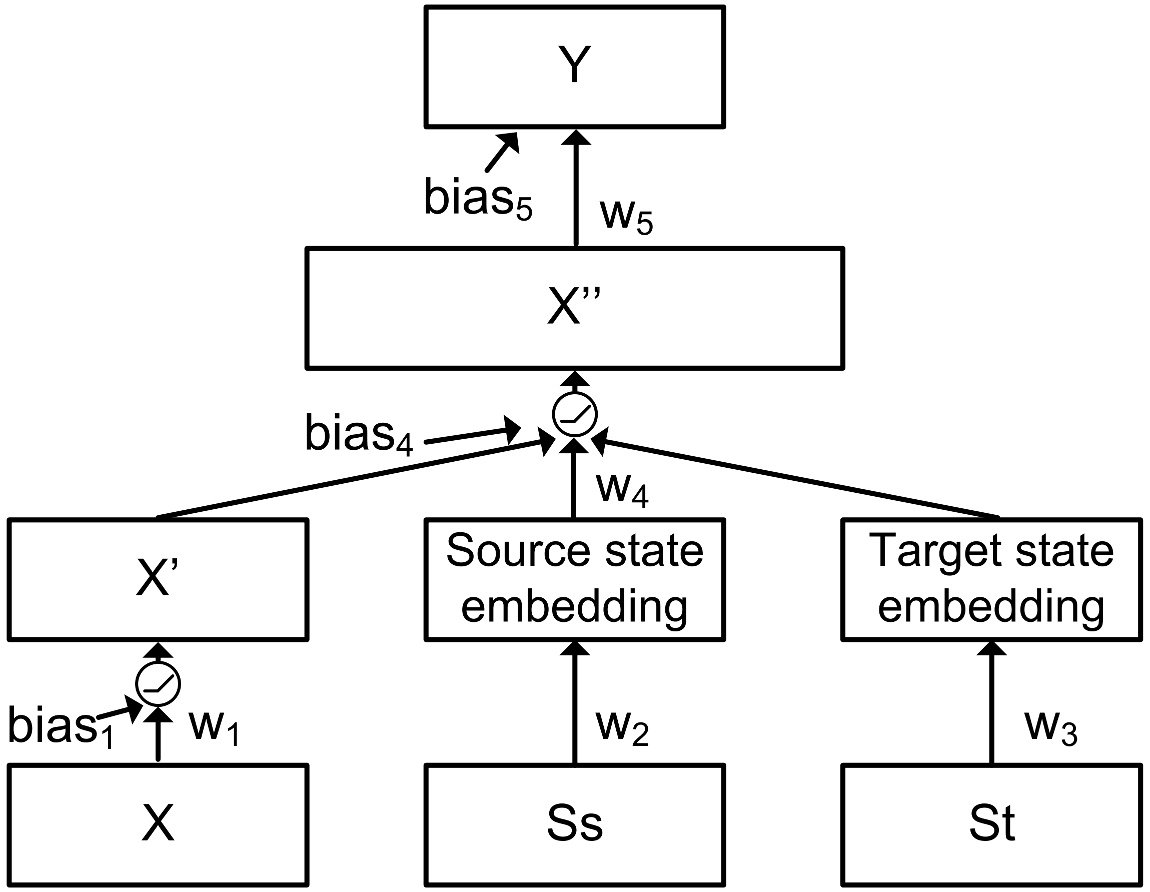
Rectified Linear Units[2]

For voice conversion we use a feed-forward ANN with rectified linear units:

$$\mathrm{RL}(x) = \max(0, x).$$

For backpropagation the gradient of $RL(x)$ is set to 0 when $x \leq 0$ and 1 when $x > 0$ ignoring discontinuity at $x = 0$

Architecture of the ANN



[2] M.D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q.V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, G. Hinton "On Rectified Linear Units for Speech Processing" *In IEEE International Conference on Acoustic Speech and Signal Processing* (ICASSP 2013) Vancouver.

6. Training the ANN (2)

To reduce oversmoothing we use variation of the speakers' state vectors

$S_s = [f_{0,s}^{min}, f_{0,s}^{max}, u_s]$ $S_t = [f_{0,t}^{min}, f_{0,t}^{max}, u_t]$

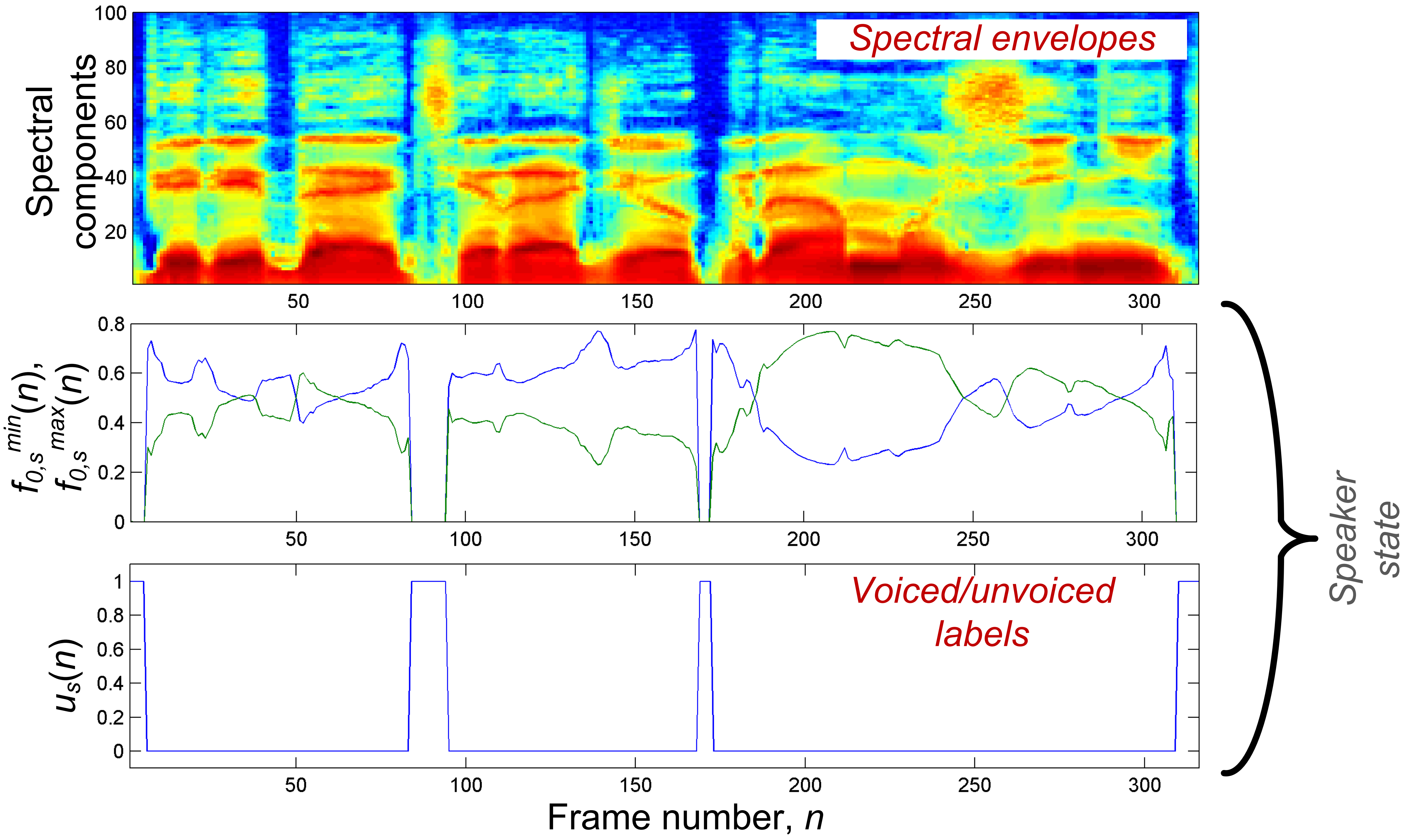
$$f_{0,x}^{min}(n) = \begin{cases} (1 - u_x(n)), & f_{0,x}(n) < F_x^{min} \\ \frac{f_{0,x}(n) - F_x^{max}}{F_x^{min} - F_x^{max}}(1 - u_x(n)), & F_x^{min} \leq f_{0,x}(n) \leq F_x^{max}, \\ 0, & f_{0,x}(n) > F_x^{max} \end{cases}$$

$$f_{0,x}^{max}(n) = \left(1 - f_{0,x}^{min}(n)\right)(1 - u_x(n)),$$

F_x^{min} and F_x^{max} – minimum/maximum allowed pitch values, $f_{0,x}(n)$ – current pitch value, $u_x(n)$ – current unvoiced flag that equals to 1 when frame n is unvoiced and 0 otherwise. Both S_s and S_t vectors contain normalized values in the range $[0,1]$.

7. Speaker state vector: example

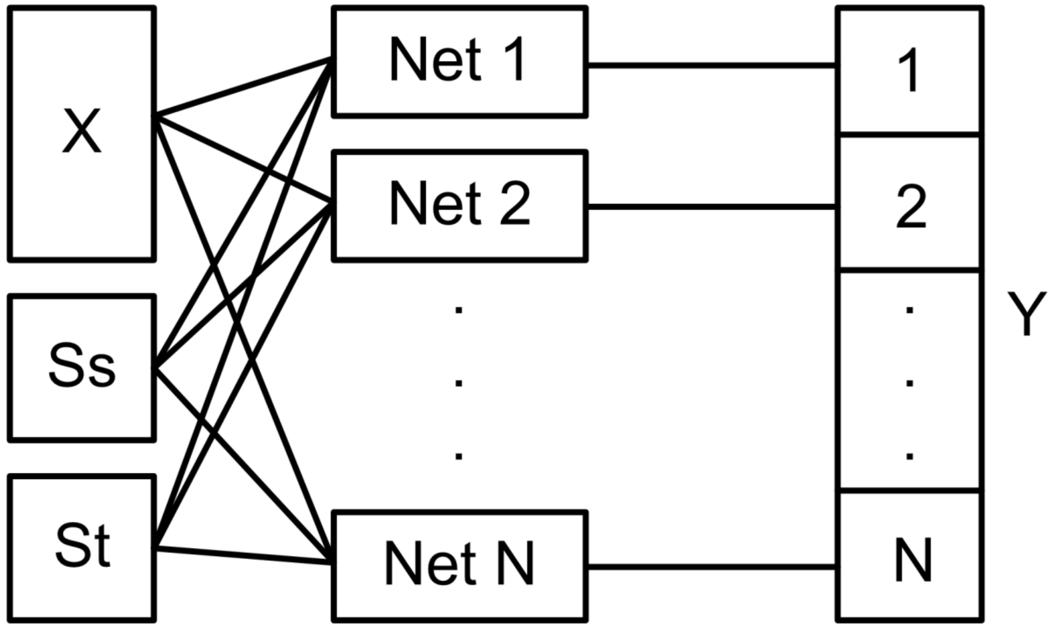
Speaker state vectors extracted from male speech.



8. Improving ANN's performance

Splitting the network

Output vector *Y* can be split into N vectors of lesser dimensionality. That simplifies the mapping task reducing it to N independent tasks. Another advantage is capacity for parallel training using a multicore processing unit.



Configuration of the network

The first hidden layer: number of neurons – 20, dimensionalities of state embeddings – 10; the second hidden layer: number of neurons – 20; dimensionality of output vectors – 5. Since target spectral envelopes are represented as 100-dimension vectors (for sample rate 44.1kHz) the number of nets N=20.

9. Real-time voice conversion

Inherent delay

The overall inherent delay of the conversion system (constant time lag between source and processed signals) is 250ms.

Computational complexity

Features are extracted with 5 ms shift that requires execution of 200 forward-propagation algorithms per second. In the most complex case (conversion 44.1 to 44.1kHz) 19.8 millions multiplications are needed.

Implementation

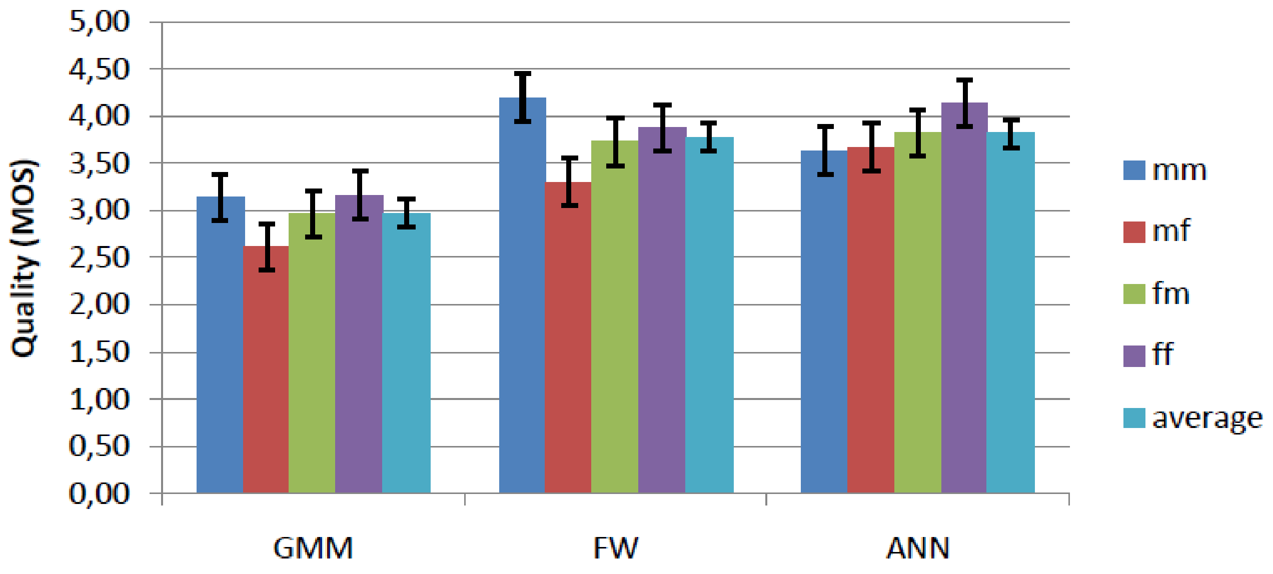
A C++ implementation of the whole real-time voice conversion system (including feature extraction and synthesis in 44.1 to 44.1 mode) has been tested on an Intel Core 2 Duo CPU (T6400 2.00 GHz using one core). The average CPU core usage is about 80%.

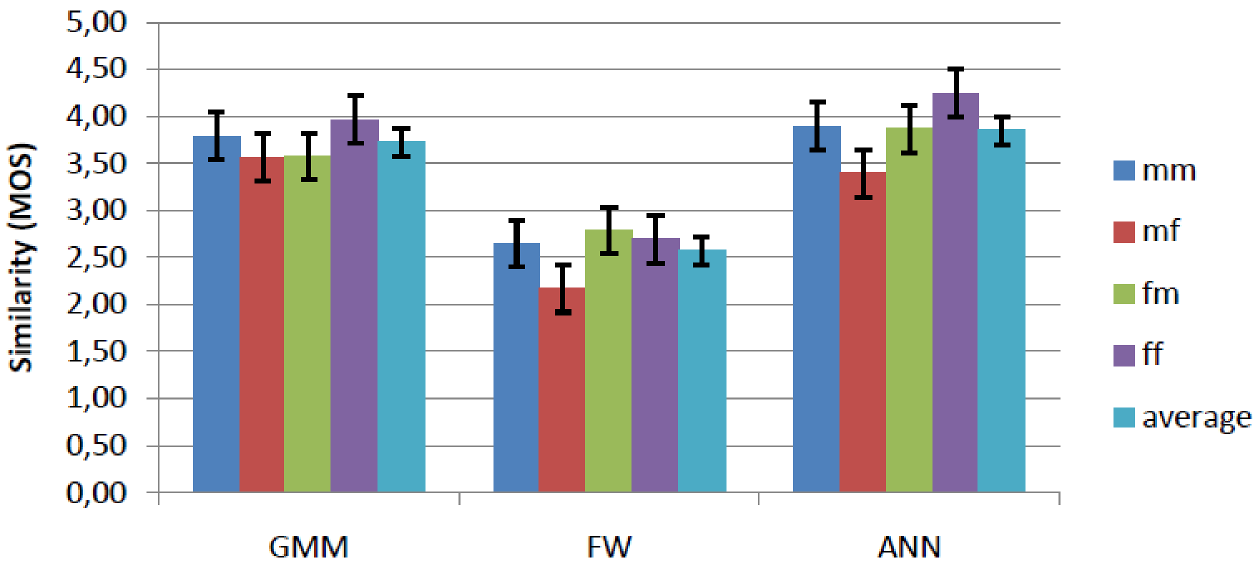
10. Subjective (MOS) evaluations

Setup

The proposed method (labeled as ‘ANN’) is compared with general GMM[3] and FW[4] methods. Training for all the methods has been made on a short (1 minute) training set.

MOS evaluations results[5]





[3] T. Toda, A.W. Black, and K. Tokuda. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *IEEE Trans. Audio, Speech and Language Processing,* Vol. 15, No. 8, pp. 2222-2235, 2007

[4] D. Erro, E. Navas, and I. Hernaez. Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. *IEEE Trans. Audio, Speech and Language Processing,* Vol. 21, No. 3, pp. 556-566, 2013.

[5] Some samples can be found on the Interspeech 2013 CDROM or on the web at http://dsp.tut.su/Package.zip

11. Conclusions

A real-time voice conversion technique based on ANN has been presented. The proposed architecture of the ANN utilizes ReLU and uses temporal speaker states in order to reduce oversmoothing effect. The spectral mapping is scalable in the sense that it allows to process signals with different sample rates. The performance of the technique has been compared with GMM and FW-based mappings using objective and subjective measures.

12. Acknowledgements

The authors are grateful to the IT-Mobile company for their support. The conversion technique presented in the paper is used in the web services 'CloneVoice' and 'CloneAudioBook'. Both services are available for users on at: http://clonevoice.com/en