

()

" + "

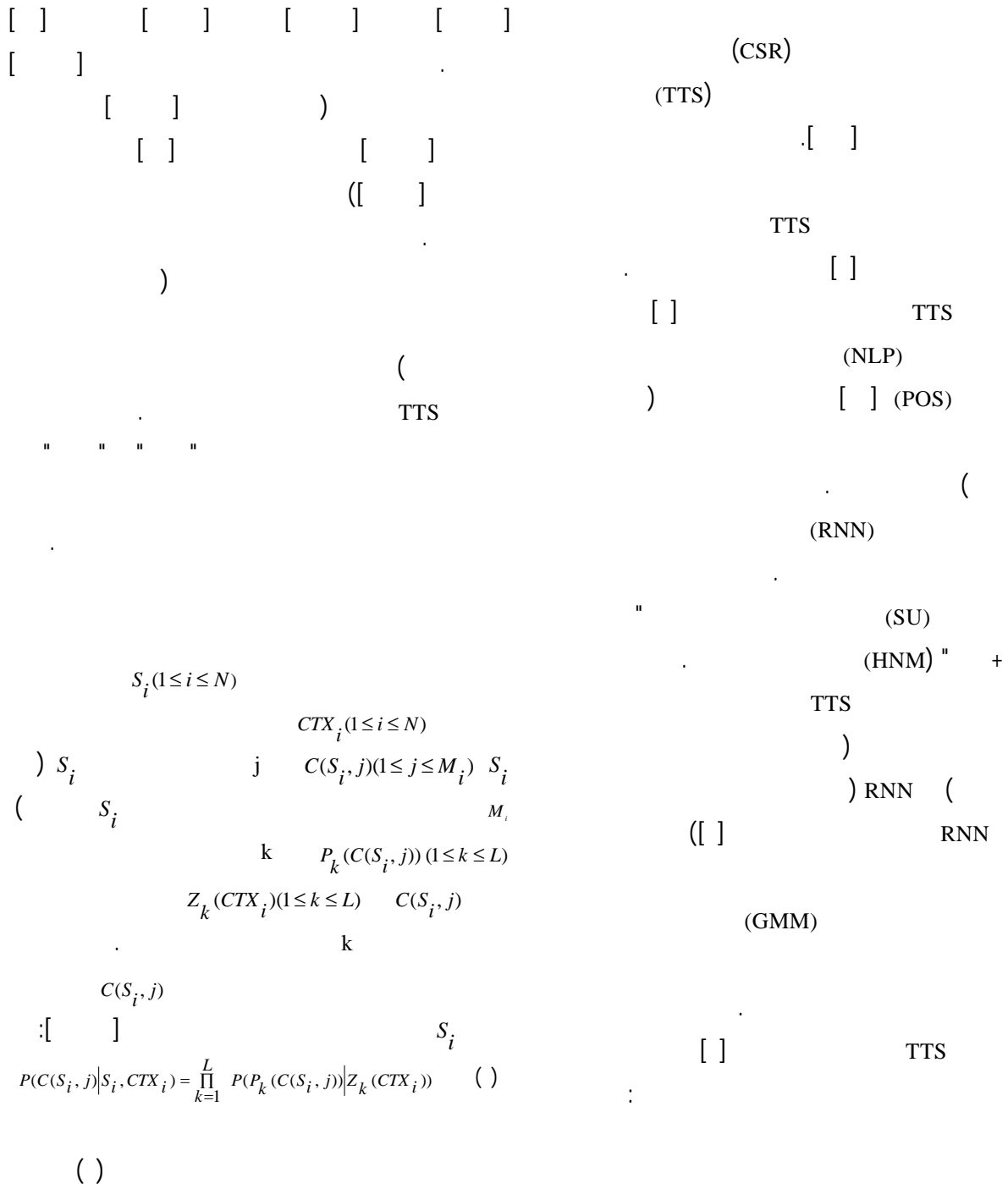
)

(

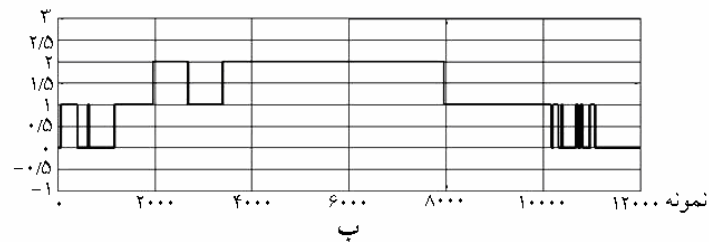
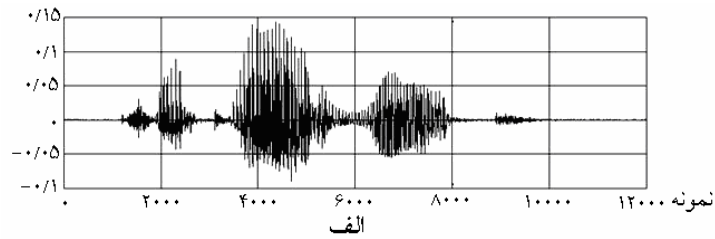
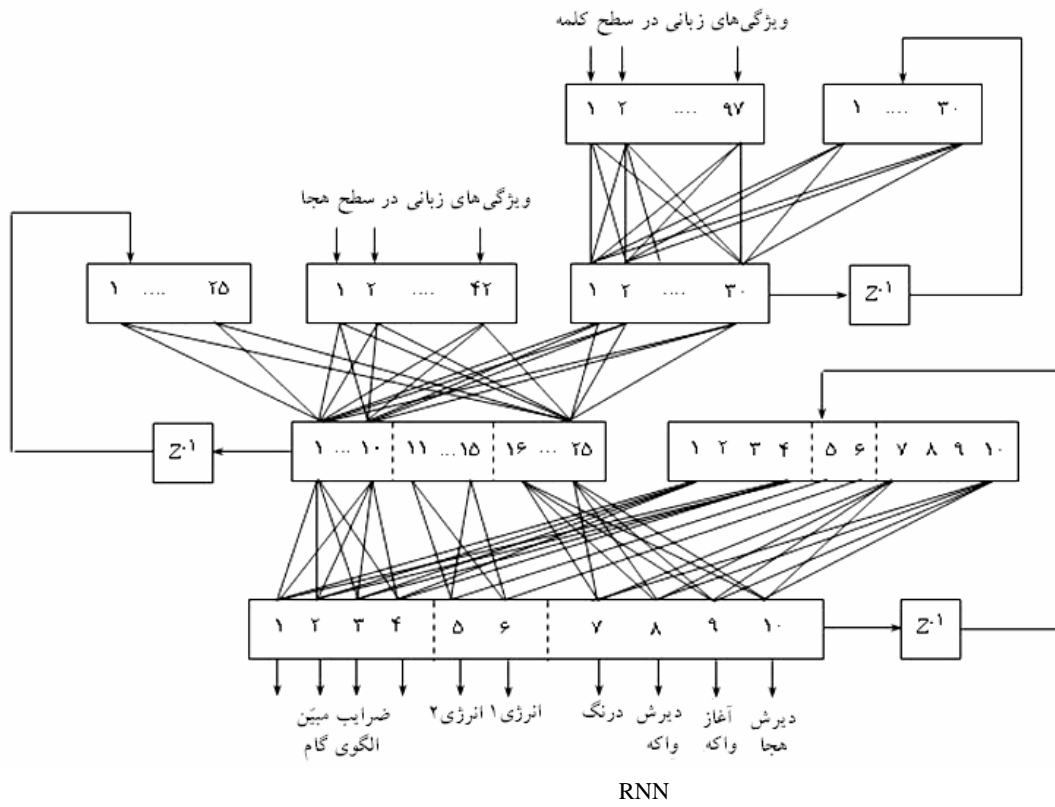
/ MOS ITU-T P.85

:

*



¹ Continuous Speech Recognition ² Text To Speech ³ Natural Language Processor
⁴ Part Of Speech ⁵ Recurrent Neural Network ⁶ Single Unit ⁷ Diphones
⁸ Harmonic + Noise Model ⁹ Prosody ¹⁰ Pitch contour ¹¹ Duration
¹² Pause ¹³ Decision trees ¹⁴ Gaussian Mixture Model ¹⁵ Rule-based
¹⁶ Data-driven ¹⁷ Hybrid ¹⁸ Target ¹⁹ Transition
²⁰ Context



V/U/S

" " "

.()"

.

:

)

[]

(

)"

"

(

$$H_i(z) = \frac{G(i)}{1 + a_1 z^{-1} + \dots + a_{13} z^{-13}}$$

()

RNN

: ()

G(i)

$$G(i) = \sqrt{\sum_{n=\langle 250 \rangle} r^2(n)}$$

()

r(n)

[]

)

AR

(m)

(

)

(σ)

(

:

()

$$T_{U/S} = m + K \cdot \sigma$$

()

K

(dB

)

/

/

log V(i)

kHz

$T_{U/S}$

msec

bit

/

)

AR

(

(" " " ")

(V/U/S)

: []

()

"

"

$$\text{شاخص انرژی} = \sqrt{\text{انرژی 1} \times \text{انرژی 2}} \quad ()$$

.(//

V/U/S)

/

)

[]

*

(/max(abs(signal)))

)

.(

/

kHz

Hz

:

:

$$V(i) = \frac{1}{N_i} \sqrt{\sum_{m=17}^{256} |H_i(e^{j\pi m/256})|^2} \quad ()$$

N_i

i

: ()

$H_i(z)$

²⁵ Event

²⁹ Auto-Regressive

²⁶ Labeling

³⁰ Volume function

²⁷ Reflection coefficient

²⁸ Residue

		G	LFV
		A	
		B	
		G	HFV
		A	
		B	
		T _{up}	
		T _{low}	
VBS	VCS		

()

()

$H_i(z)$. ()

(HFV) (LFV)

: B A G

$$R(i) = \frac{LFV(i)}{HFV(i)} \quad ()$$

R(i)

()

$$Score(i) = \begin{cases} 1 & ; R(i) \geq T_{up} \\ 0 & ; R(i) < T_{low} \\ \frac{R(i) - T_{low}}{T_{up} - T_{low}} & ; T_{low} \leq R(i) < T_{up} \end{cases} \quad ()$$

(MLP)

()

MLP / () Hz

kHz kHz

[]

()

³¹ Sonorant ³² Voiced consonant ³³ Nasal ³⁴ Semivowel

³⁵ Voiced fricative ³⁶ Low Frequency Volume function ³⁷ High Frequency Volume function ³⁸ Voice bar

³⁹ Multi-Layer Perceptron

$p_n p_n p_n$ " + " .
 $p_n p_n$ ($d_n d_n d_n p_n$)
) RNN ()
 " " " " () RNN
 (e_n, e_n) " "
 n : ()
 / kHz
 kHz / kHz
 $n- n$ ($p_n p_n, \dots, p_n p_n$)
 msec
 ($d_n d_n, \dots, d_n d_n$) " "
 % /
 " " ($e_n e_n, e_n e_n$)
 " " " (CVCC CVC CV)

$$P_{GMM}(Z; \alpha, \mu, \Sigma) = \sum_{q=1}^Q \alpha_q N(Z; \mu_q, \Sigma_q) \quad ()$$

$$\sum_{q=1}^Q \alpha_q = 1 (\alpha_q \geq 0)$$

$$N(Z; \mu, \Sigma) . \quad q$$

$$N(Z; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} \sqrt{|\Sigma|}} \exp\left[-\frac{1}{2}(Z-\mu)^T \Sigma^{-1} (Z-\mu)\right] \quad ()$$

[] :
 . [] []

{ α, μ, Σ } GMM

()

[]

EM

GMM

) []

(HNM

GMM

GMM

[]

)

[]

$P_n P_n$)

(

GMM

($P_n P_n, P_n P_n, P_n P_n,$

HNM1

(Q)

[]

[] (MVF)

GMM

)

HNM

[]

(

[]

[]

[]

GMM

[]

GMM

RNN

(UV)

GMM

[]

)

RNN

(

RNN

⁴¹ Expectation – Maximization
⁴⁵ Smoothing

⁴² Diagonal

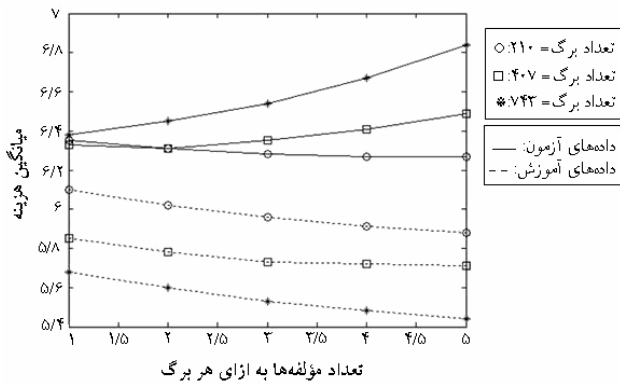
⁴³ Concatenative synthesis

⁴⁴ Maximum Voiced Frequency

RNN		RMSE		
(msec)	(msec)	(msec)	(dB)	(Hz)
/	/	/	/	/



" " RNN



[] ITU-T P.85

: (MOS)
 ()
 ()
 ()
 ()

" " MATLAB

RMSE
 RNN

()

"

)

(RNN

RNN

GMM

()

"

⁴⁶ Toolbox
⁵⁰ Likelihood
⁵⁴ Listening effort
⁵⁸ Pleasantness

⁴⁷ MATrix Laboratory
⁵¹ Overtraining
⁵⁵ Comprehension

⁴⁸ Root Mean Squared Error
⁵² Mean Opinion Score
⁵⁶ Articulation

⁴⁹ Cost
⁵³ Overall impression
⁵⁷ Pronunciation

(FTTSv.1 FTTSv.2)

TTS

(FTTSv.1 FTTSv.2)								TTS
FTTSv.2	FTTSv.1	ATT	SS	RS	AK	LT	EL	
/	/	/	/	/	/	/	/	
/	/	/	/	/	/	/	/	
/	/	/	/	/	/	/	/	
/	/	/	/	/	/	/	/	
/	/	/	/	/	/	/	/	
/	/	/	/	/	/	/	/	
/	/	/	/	/	/	/	/	()

() () ()
 ()

)

(

FTTSv.2)

NLP

TTS

(

FTTSv.1

GMM

SS RS AK LT EL

GMM

) ATT

([] ITU-T P.85

FTTSv.1

[]

HNM

RNN

RNN

RMSE

GMM

[]

ITU-T P.85

()

/ MOS

)

⁵⁹ Elan Informatique
⁶³ Speechworks Speechify

⁶⁰ Lucent Technology
⁶⁴ American Telephone and Telegraph

⁶¹ Aculab

⁶² RealSpeak

[]

- [10] Sheikhan M., Tebyani M., Lotfizad M., Using symbolic and connectionist approaches to automate editing Persian sentences syntactically; Proceedings of International Conference on Intelligent and Cognitive Systems 1996; 250-253.
- [11] Yamashita Y., Ishida T., Stochastic F0 contour model based on the clustering of F0 shapes of a syntactic unit; Proceedings of European Conference on Speech Communication and Technology 2001; 533-537. []
- [12] Buhmann J., Marten J.P., Macken L., Van Coile B., Intonation modeling for the synthesis of structured documents; Proceedings of International Conference on Spoken Language Processing 2002; 2089-2092.
- [13] Payo V.C., Mancebo D.E., A strategy to solve data scarcity problems in corpus based intonation modeling; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2004; 665-668. []
- [14] Agüero P.D., Bonafonte A., Consistent estimation of Fujisaki's intonation model parameters; Proceedings of International Conference on Speech and Computer 2005; 297-300.
- [15] Ishi C.T., Ishiguro H., Hagita N., Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality; Speech Communication 2008; 50:531-543.
- [16] Smith C.L., Modeling durational variability in reading aloud a connected text; Proceedings of International Conference on Spoken Language Processing 2002; 1769-1772.
- [17] Eichner M., Wolff M., Hoffmann R., Improved duration control for speech synthesis using a multigram language model; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2002; 417-420. []
- [18] Werner S., Wolff M., Eichner M., Hoffmann R.; Modeling pronunciation variation for spontaneous speech synthesis; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2004; 673-676. []
- [19] Ariu M., Masuko T., Tanaka S., Kawamura A., Speech recognition using syllable duration ratio model; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2006; 341-344. []
- [20] Sagisaka Y., Sato H.; Accentuation rules in Japanese TTS conversion; Rev. Elect. Commun. Lab. 1984; 32:188-199. : ()
- [21] Low P.H., Vaseghi S., Application of microprosody models in TTS synthesis; Proceedings of International Conference on Spoken Language Processing 2002; 2413-2416. []
- [22] Hifny Y., Rashwan M., Duration modeling for Arabic TTS synthesis; Proceedings of International Conference on Spoken Language Processing 2002; 1773-1776. []
- [3] Sheikhan M., Tebyani M., Lotfizad M., Continuous speech recognition and syntactic processing in Iranian Farsi language; International Journal of Speech Technology 1997; 1:135-141. []

- [39] Wutiwathchai C., Furui S., Thai speech processing technology: a review; *Speech Communication* 2007; 49:8-27.
- [40] Chen K., Hasegawa-Johnson M., Cohen A., An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model; *Proceedings of International Conference on Acoustics, Speech, and Signal Processing* 2004; 509-512.
- [41] Ma X., Zhang W., Zhu W., Shi Q., Jin L., Probability based prosody model for unit selection; *Proceedings of International Conference on Acoustics, Speech, and Signal Processing* 2004; 649-652.
- [42] Sun X., Applebaum T.H., Intonational phrase break prediction using decision tree and N-gram model; *Proceedings of International Conference on Acoustics, Speech, and Signal Processing* 2000; 1281-1284.
- [43] Jiang D., Shi Q., Meng F., Shuang Z., Ma X., Lin Y., Overview of the IBM Mandarin text-to-speech system; *Proceedings of Technology and Corpora Workshop on Speech-to-Speech Translation* 2006; 181-185.
- [44] Bahl L.R., de Souza P.V., Gopalakrishnan P.S., Nahamoo D., Decision tree for phonological rules in continuous speech; *Proceedings of International Conference on Acoustics, Speech, and Signal Processing* 1991; 185-188.
- [45] Sheikhan M.; RNN-based prosodic information synthesizer for Farsi TTS; *Second Irano-Armenian Workshop on Neural Networks*, Dec. 1999.
- [46] Chen S.H., Wang Y.R., Vector quantization of pitch information in Mandarin speech; *IEEE Trans. Commun.* 1990; 38:1317-1320.
- [47] Childers D.G., Time modification of speech, theory: speech analysis, segmentation and labeling, In *Speech Processing and Synthesis Toolboxes*; John Wiley & Sons, Inc.; 2000:330-359.
- []
- ()
- [49] Ghahramani Z., Jordan M.I., Supervised learning from incomplete data via an EM approach, In *Advances in Neural Information Processing Systems*; MIT Press; 1994:120-127.
- [50] Weiss B., Prosodic elements of a political speech and its effects on listeners; *Proceedings of International Conference on Speech and Computer* 2005; 127-130.
- [51] Moulines E., Emerard F., Larreur D., Le Saint Milon J.L., Le Fancheur L., Marty F., Charpentier F., Sorin C., A real-time French TTS system generating high quality synthetic speech; *Proceedings of International Conference on Acoustics, Speech, and Signal Processing* 1990; 309-312.
- [52] Bulut M., Narayanan S.S., Syrdal A.K., Expressive speech synthesis using a concatenative synthesizer; *Proceedings of International Conference on Spoken Language Processing* 2002; 1265-1268.
- [53] Erro D., Moreno A., A pitch-asynchronous simple method for speech synthesis by diphone concatenation using the deterministic plus stochastic model; *Proceedings of International Conference on Speech and Computer* 2005; 321-324.
- [23] El-Imam Y.A., Synthesis of the intonation of neutrally spoken modern standard Arabic speech; *Signal Processing* 2008; 88:2206-2221.
- [24] Frid J., Prediction of intonation patterns of accented words in a corpus of read Swedish news through pitch contour stylization; *Proceedings of European Conference on Speech Communication and Technology* 2001; 915-918.
- [25] Lobanov B., Tsurilnik L., Zhadinets D., Piorokovska B., Rafalko J., Szpilevsky E., Language-specific application of intonation contours in Russian and Polish multilingual TTS synthesis; *Proceedings of International Conference on Speech and Computer* 2005; 317-320.
- [26] Kaiki N., Mimura K., Sagisaka Y., Statistical modeling of segmental duration and power control for Japanese; *Proceedings of European Conference on Speech Communication and Technology* 1991; 625-628.
- [27] Fukuda T., Komori Y., Aso T., Ohora Y., A study of pitch pattern generation using HMM-based statistical information; *Proceedings of International Conference on Spoken Language Processing* 1994; 723-726.
- [28] Fujio S., Sagisaka Y., Higuchi N.Z., Stochastic modeling of pause insertion using context-free grammar; *Proceedings of International Conference on Acoustics, Speech, and Signal Processing* 1995; 604-607.
- [29] Taylor P., Black A.W., Assigning phrase breaks from part-of-speech sequences; *Computer Speech and Language* 1998; 12:99-117.
- [30] Bulyko I., Ostendorf M., Joint prosody prediction and unit selection for concatenative speech synthesis; *Proceedings of International Conference on Acoustics, Speech, and Signal Processing* 2001; 781-784.
- [31] Adell J., Agüero P.D., Bonafonte A., Database pruning for unsupervised building of text-to-speech voices; *Proceedings of International Conference on Acoustics, Speech, and Signal Processing* 2006; 889-892.
- [32] Sakai S., Shu H., A probabilistic approach to unit selection for corpus-based speech synthesis; *Proceedings of International Speech Communication Association Conference* 2005; 81-84.
- [33] Scordilis M.S., Gowdy J.N., Neural network-based generation of fundamental frequency contours; *Proceedings of International Conference on Acoustics, Speech, and Signal Processing* 1989; 219-222.
- [34] Taylor P., Using neural networks to locate pitch accents; *Proceedings of European Conference on Speech Communication and Technology* 1995; 1345-1348.
- [35] Riedi M., A neural-network-based model of segmental duration for speech synthesis; *Proceedings of European Conference on Speech Communication and Technology* 1995; 599-602.
- [36] Callan D., Tajima K., Callan A., Akahane-Yamada R., Masaki S., Neural processes underlying perceptual learning of difficult second language phonetic contrast; *Proceedings of European Conference on Speech Communication and Technology* 2001; 145-148.
- [37] Teixeira J.P., Freitas D.; Use of phoneme dedicated artificial neural networks to predict segmental durations; *Proceedings of International Conference on Speech and Computer* 2005; 679-682.
- [38] Espinosa H.P., Reyes Garcia C.A., Genetic algorithms for the selection of cellular automata rules applied to intonation modeling in text to speech synthesis; *Proceedings of International Conference on Speech and Computer* 2005; 671-674.

-
- [60] Zeljkovic I., Stylianou Y., Single complex sinusoid and ARHE model based pitch extractors; Proceedings of European Conference on Speech Communication and Technology 1999; 2813-2816.
- [61] Stylianou Y., Removing linear phase mismatches in concatenative speech synthesis; IEEE Trans. Speech Audio Processing 2001; 9:232-239.
- [62] Quatieri T.F., McAulay R.J., Shape invariant time-scale and pitch modification of speech; IEEE Trans. Signal Processing 1992; 40:497-510.
- [63] Dutiot T., Stylianou Y., Text-to-speech synthesis, In: Oxford Handbook of Computational Linguistics; Oxford University Press; 2003:323-338.
- [64] O'Brien D., Monaghan A., Shape invariant time-scale modification of speech using a harmonic model; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 1999; 381-384.
- [65] ITU-T Recommendation P. 85: A method for subjective performance assessment of the quality of speech output devices; International Telecommunications Union Publication; 1994.
- [66] Alvarez Y.V., Huckvale M., The reliability of the ITU-T P. 85 standard for the evaluation of TTS systems; Proceedings of International Conference on Spoken Language Processing 2002; 329-332.
- [54] Toda T., Kawai H., Tsuzaki M., Optimizing sub-cost functions for segment selection based on perceptual evaluations in concatenative speech synthesis; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2004; 657-660.
- [55] Nukaga N., Komoshida R., Nagamatsu K., Kitahara Y., Scalable implementation of unit selection based text-to-speech system for embedded solutions; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2006; 849-852.
- [56] Storm V., Clark R., King S., Expressive prosody for unit-selection speech synthesis; Proceedings of International Speech Communication Association Conference 2006; 1296-1299.
- [57] Conkie A., Syrdal A.K., Expanding phonetic coverage in unit selection synthesis through unit substitution from a donor voice; Proceedings of International Speech Communication Association Conference 2006; 1754-1757.
- [58] Mohammadi M., Sheikhan M., TTS in broadcasting; Proc. International Conference BroadcastAsia 2000; 35-37.
- [59] O'Brien D., Monaghan A., Concatenative synthesis based on a harmonic model; IEEE Trans. Speech Audio Processing 2001; 9:11-20.