I.J. Information Technology and Computer Science, **2013, 04, 23-29**
Published Online March 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2013.04.03

Evaluation of Hidden Semi-Markov Models Training Methods for Greek Emotional Text-to-Speech Synthesis

Alexandros Lazaridis[1], Iosif Mporas[1,2]
[1]Dept. of Electrical and Computer Engineering, University of Patras, Rion-Patras 26500, Greece
[2]Dept. of Informatics and Mass Media, Technological Educational Institute of Patras, Pyrgos 27100, Greece
Emails: {alaza, imporas}@upatras.gr

Abstract— This paper describes and evaluates four different HSMM (hidden semi-Markov model) training methods for HMM-based synthesis of emotional speech. The first method, called emotion-dependent modelling, uses individual models trained for each emotion separately. In the second method, emotion adaptation modelling, at first a model is trained using neutral speech, and thereafter adaptation is performed to each emotion of the database. The third method, emotion-independent approach, is based on an average emotion model which is initially trained using data from all the emotions of the speech database. Consequently, an adaptive model is build for each emotion. In the fourth method, emotion adaptive training, the average emotion model is trained with simultaneously normalization of the output and state duration distributions. To evaluate these training methods, a Modern Greek speech database which consists of four categories of speech, anger, fear, joy and sadness, was used. Finally, an emotion recognition rate subjective test was performed in order to measure and compare the ability of each of the four approaches in synthesizing emotional speech. The evaluation results showed that the emotion adaptive training achieved the highest emotion recognition rates among four evaluated methods, throughout all four emotions of the database.

Index Terms— HMM Synthesis; Emotional Synthesis; HSMM Adaptation

I. Introduction

Over the last years unit selection corpus-based technique [1-4] is widely used approach in the field of speech synthesis. This method is mainly based on clustering units of a large speech database according to distance criteria and selecting the appropriate ones during runtime according to matching criteria [1]. Unit selection corpus-based speech synthesis systems are shown to produce synthetic speech of high naturalness and intelligibility, concatenating speech units with no or almost no signal processing of them. In this technique, the style characteristics of the synthetic speech follow the ones of the recorded speech of the database. This leads to limitations in the variation of speaking styles, emotions or voice characteristics of the synthetic speech since each time large databases need to be recorded following these variations or styles. This drawback of the unit selection speech synthesis approach led to the development of the statistical parametric speech synthesis and mainly to the hidden Markov model (HMM)-based speech synthesis [5-9].

In HMM-based synthesis, HMMs are trained using speech databases of natural speech and the synthetic speech is generated by them through the Mel log spectral approximation (MLSA) filter [10]. This technique, in contrast to the unit selection corpus-based approach, even though it produces synthetic speech of lower quality, offers the advantage of modelling different speaking styles and emotions with the use of limited databases. This is achieved by applying a model adaptation algorithm, such as the Maximum Likelihood Linear Regression (MLLR) algorithm [11,12] or MAP-based (Maximum A Posteriori) modification [13-15], using a small amount of speech uttered by the target speaker. The target speaker is not restricted only to different speakers with different voice characteristics but also can be characterized by different speaking styles or even different emotions.

Over the last decades, the increasing interest in human-computer interaction and spoken dialog systems raised the need for more effective and user-friendly systems. In this direction, in the field of speech synthesis, the implementation of various speaking styles and emotions becomes more and more important, making HMM-based speech synthesis more and more appropriate for emotional speech synthesis. In this paper, a comparison evaluation on HSMM (hidden semi-Markov model) training approaches in HMM-based speech synthesis for synthesizing speech of various emotions is presented. Specifically, four different HSMM training approaches are used for synthesizing emotional speech. In the first approach, called emotion-dependent modelling, the speech data corresponding to each emotion separately is used in order to train the model producing synthetic speech of this emotion. In the second approach, called emotion adaptation modelling, initially a database of neutral speech is used for training a model. Consequently this model in adapted to the target emotion with a MLLR-based adaptation technique using the speech data corresponding to each emotion separately in order to synthesize emotional speech of the specific emotion. In the third approach, called emotion-independent modelling, an initial model, the average emotion model, is trained using the speech data of all the emotional categories of the database. In this method a MLLR-based adaptation technique is used for adapting the model created using all the speech data, to the target emotion. Finally a method called emotion adaptive training, based on speaker adaptive training approach (SAT) [16,17], was implemented in our experiments. This technique is based on the implementation of the average emotion model and the simultaneously normalization of its output and state duration distributions.

A speech database of emotional speech of Modern Greek, consisting of four categories of emotional speech: anger, fear, joy and sadness, along with a second database of neutral speech of Modern Greek are used in the experiments. A subjective test measuring the emotion recognition rate was performed evaluating the ability of each approach to synthesize emotional speech.

The rest of the paper is organized as follows. In Section II, a description of the HSMM-based synthesis system, along with the four methods evaluated in this work, is given. In Section II, the speech databases used in our experiments along with the experimental protocol are presented. The experimental results are presented in Section IV. Finally, the conclusions of this work are reported in Section V.

II. Hidden Semi-Markov Model-based Speech Synthesis

In HMM-based parametric speech synthesis, the spectrum, pitch and duration of natural speech are simultaneously modelled in a unified framework of HMMs [6]. Specifically, in the training phase, the spectrum is modelled by continuous probability distribution HMMs, the pitch is modelled by multi-space probability distribution HMMs and the state durations by multi-space Gaussian distribution. Multi-space distribution models are used in order to overcome the problem of non continuous pitch values in the parts of unvoiced speech [18]. Context-dependent decisions trees are used for clustering independently the distributions of the spectral, pitch and duration parameters. In synthesis phase, a sentence HMM is created by concatenating phoneme HMMs based on the input text. Consequently speech parameters vector sequences are generated by the concatenated phoneme HMMs using a speech parameter generation algorithm [19]. Finally the synthetic speech is generated through the MLSA filter [10].

Nonetheless, an inconsistency is raised in this procedure since state duration models are explicitly used during the synthesis phase without being incorporated in the training phase. This drawback of HMMs can be overcome with the use of the HSMMs [20]. HSMMs are characterized by their ability to incorporate the explicit modelling of state durations not only in the synthesis phase as HMMs do, but also in the training phase of the HSMM-based speech synthesis systems improving the naturalness of synthetic speech [20]. In the following subsections four approaches for HSMM modelling of emotional speech are described.

2.1 Emotion-Dependent (ED) Method

In this method, called emotion-dependent (ED) modelling, each emotion is individually modelled by an acoustic model using only the data corresponding to this emotion [21]. An additional root node is implemented in the clustering decision tree having as leaves the corresponding decision tree of the respective emotion. The training and synthesis phases follow the respective procedures described above in Section II.

2.2 Emotion Adaptation (EA) Method

In emotion adaptation (EA) method, in the training phase, a neutral speech database is used to train an initial model following the respective procedure described earlier in Section II. In the adaptation phase, the data corresponding to the target emotion are used to adapt the initial neutral model to the model of the target emotion. Specifically, an MLLR adaptation [12] is applied, transforming both the output and state duration distributions of the HSMMs. In this adaptation technique two categories of regression matrices are created, one for the output distribution and one for the state duration distribution, so as to maximize the likelihood of the adaptation data [12]. In synthesis phase, the adapted to the target emotion model is used to produce synthetic speech following the respective procedure described above in Section II.

2.3 Emotion-Independent (EI) Method

In the emotion-independent (EI) method, an average emotion model is trained, respectively to the average voice model [22,23], using a multi-emotional database. Consequently this model is adapted to the target emotion using the respective data. Specifically, in the training phase, the emotion-dependent models, one for each emotion of the database, apart from the target emotion, are firstly separately trained using the multi-emotional speech database. Consequently these context-dependent emotion models are clustered using a shared decision tree, creating an emotion-independent decision tree. The average emotion model is created by combining, at each leaf node of the decision tree, Gaussian probability distribution functions of the emotion-dependent models [22,23]. In order for all the nodes of the decision tree to have data from all the emotions, during the split of a node of the decision tree, only the context related questions which can split the node for all emotion-dependent models, are used. The adaptation of the average model to the target emotion is achieved through a MLLR adaptation [12] transformation of the output and state duration distributions as described in Section 2.2. In synthesis phase the HMMs generate speech parameters vector sequences through the speech parameter generation algorithm [19] and finally the synthetic speech is generated through the MLSA filter [10].

2.4 Emotion Adaptive Training (EAT) Method

The emotion adaptive training (EAT) technique, respectively to the speaker adaptive training (SAT) [16,17], normalizes simultaneously the output and state duration distributions of the average emotion model. In EAT, the MLLR adaptation is used as an emotion normalization technique of the average emotion model to reduce the influence of emotion differences and acoustic variability of spectral and pitch parameters [23]. Basically, in the HSMM-based EAT approach the parameter set of HSMM and the set of transformation matrices, for each training emotion in respect to the average emotion model, are estimated simultaneously [23] maximizing the likelihood of the training data. After the average emotion model is trained the adaptation of the model to the target emotion is achieved using MLLR approach [12] as described in Section 2.2 and the synthetic speech is generated through the MLSA filter [10], as described previously in Section II.

III. Experimental Setup

3.1 Speech Databases

Two Modern Greek speech databases were used in our experiments, one emotional speech database containing five emotional categories [24] and one database containing only neutral speech [25].

3.1.1 Emotional Speech Database

The content of the emotional speech database was extracted from passages, newspapers or were set up by a professional linguist. This database is linguistically and prosodically rich, and contains emotional speech from the categories: anger, fear, joy, sadness, which are considered as the four archetypal emotions [26], as well as neutral speech. The database consisted of 62 utterances, which were pronounced several times with different emotional charge. The length of the utterances was ranging from a single word, a phrase, short and long sentence or even a sequence of sentences of fluent speech. The context of all sentences was emotionally neutral, meaning that it did not convey any emotional charge through lexical, syntactic or semantic means. Moreover, all the utterances were uttered separately in the five emotional styles. The entire database consisted of 4150 words (310 utterances). All utterances were uttered by a professional, female actress, speaking Modern Greek. To ensure that the speaker would not have to change her emotional state more than five times, expressing anger, fear, joy, neutral and sadness emotion respectively, all the recordings of each specific emotional category were recorded in series, before proceeding with the other emotional categories. In addition, the actress was instructed to express a 'casual' intensity of the chosen emotional state avoiding any theatrical exaggeration. All recording sessions were held in the anechoic chamber of a professional studio.

Table 1: Structural information of the emotional speech database

# Words per Sentence	Frequency of Occurence (%)
<3	14.12
4	03.53
5	08.24
6	09.41
7	07.06
8	04.71
9	04.71
10	03.53
11	04.71
12	12.94
13	07.06
14	04.71
15	01.18
16	01.18
17	01.18
>18	11.76
Total	100.00

Table 2: The twenty most frequently occurred words in the emotional speech database

Word	Pronunciation	# of Occurrence
του	tu	33
και	Ke	30
το	to	26
την	tin	22
με	me	18
για	Ya	14
της	tis	14
από	apo	12
οι	i	12
τα	ta	12
τους	tus	11
δεν	Den	10
η	i	10
στο	sto	10
τις	tis	10
τη	ti	9
στην	stin	8
να	na	7
που	pu	7
στα	sta	6

Table 1 and Table 2 show structural information of the emotional speech database. In particular, in Table 1 the number of words per sentence is presented. In Table 2 the twenty most frequent words of the database are presented along with the number of their occurrences and the pronunciation of the words. We used a phone inventory of 34 phones, with total of 22045 instances (15667 voiced and 6378 unvoiced phones). Furthermore, each vowel class included both stressed and unstressed cases of the corresponding vowel.

After recording the emotional speech database, a listening test was performed to validate the emotions of the database. Specifically, six listeners, of different ages with no particular knowledge in speech synthesis, were asked to identify the emotion that characterized each recorded utterance [24]. Five sentences were selected with all the respective emotions (twenty recordings) and played randomly to each listener. In the first part of the test, a free response was given by the listener labelling each recording with whatever emotion found appropriate, and in the second part, forced response test, the listener was classifying each recording to one of the four emotional categories included in the database (anger, fear, joy and sadness) [24]. The results of the validation subjective test are shown in Table 3. In our experiments only the four emotional categories, anger, fear, joy and sadness, were used while for neutral speech a second database was used.

Table 3: Subjective validation on the emotions of the database: free and force response tests

Emotional Categories	Free Response Test	Forced Response Test
Anger	97.8%	98.2%
Fear	68.0%	74.0%
Joy	84.0%	89.0%
Sadness	97.1%	97.5%

3.1.2 Neutral Speech Database

The second database, Vergina speech database [25], was developed in support of research and development of corpus-based unit selection and statistical parametric speech synthesis systems for the Modern Greek language. A text corpus of approximately 5 million words, collected from newspaper articles, periodicals and paragraphs of literature, was processed in order to select the utterances-sentences needed for making the speech database and to achieve a reasonable phonetic coverage. The broad coverage and contents of the selected utterances-sentences of the database -text corpus collected from different domains and writing styles- makes this database appropriate for various application domains. The database, recorded in audio studio, consists of approximately 3,000 phonetically balanced Modern Greek utterances corresponding to approximately four hours of speech. Annotation of the Vergina speech database was performed using task-specific tools, which are based on a hidden Markov model (HMM) segmentation method [27], and then manual inspection and corrections were performed. Finally it should be mentioned that in both databases, speech was sampled at 44.1 kHz, and a resolution of 16 bit.

Table 4: Structural information of Vergina speech database

# Words per Sentence	Frequency of Occurence (%)
3	00.14
4	00.48
5	06.23
6	14.27
7	20.26
8	21.96
9	21.69
10	06.26
11	03.98
12	02.45
13	00.92
14	00.65
15	00.48
16	00.14
17	00.07
18	00.03
Total	100.00

Table 5: The twenty most frequently occurred words in Vergina speech database along with their pronunciation.

Word	Pronunciation	# of Occurrence
και	Ke	586
το	to	566
να	na	545
η	i	502
του	tu	470
είναι	Ine	428
ο	o	369
την	tin	315
της	tis	311
δεν	Den	300
τα	ta	279
για	Ya	258
οι	i	249
θα	Qa	245
με	me	227
από	apO	225
σε	se	180
στο	sto	179
που	pu	171
των	ton	167

Table 4 and Table 5 show structural information of the Vergina speech database. In particular, in Figure 2 the number of words per sentence is presented. In Table

5 the twenty most frequent words of the database are presented along with the number of their occurrences and the pronunciation of the words. The phone-set used in the database is a modification of the SAMPA [28] phonetic alphabet for Greek. The phone-set consisted of 39 phones plus the silent (pau) was adopted.

3.2 Experimental Protocol

The implementation of our HSMM-based speech synthesis system is based on the HTS framework [29]. In our experiments, the speech signals of both databases were down-sampled to the frequency of 16 kHz and a phone set of 34 phones was adopted for building the HSMM-based speech synthesis systems. We used 5-state left-to-right with no-skip HSMMs. The parameters were extracted using a 25 msec Hamming-windowed frame length with a 5 msec frame shift. The feature vector consisted of 25 Mel-frequency cepstral coefficients (MFCCs), including zeroth coefficient, and logarithm of fundamental frequency (logF0). Moreover, both dynamic (delta) and acceleration (delta-delta) coefficients were used both for spectrum (MFCC) and pith (logF0) representation.

In the case of ED modelling (Section 2.1) each model was trained using 55 sentences of the respective emotion and the evaluation of them was done synthesizing the rest 7 sentences. Concerning the EA modelling (Section 2.2), 1200 sentences of Vergina speech database were used for training the initial neutral model. For the adaptation of the model, the same 55 sentences with the ones used in the ED modelling were used for each emotion and the rest ones (7 sentences) for the evaluation of the approaches. In the case of EI modelling, the 1200 sentences of neutral selected from Vergina speech database along with the same 55 sentences mentioned above for each emotion apart from the ones of the target emotion were used for training the EI model. The 55 sentences of the target emotion were used for the adaptation of the average emotion model created by the training phase of the EI modelling and the rest ones (7 sentences) of the target emotion were used for the evaluation of the model. The same sets of sentences were also used for the average emotion model created by the EAT approach.

IV. Experimental Results

In order to evaluate the effectiveness of the training approaches in producing synthetic speech of specific emotions a subjective test was performed. Ten males and two females, which were the subjects of the emotion classification test, were asked to classify the synthesized utterances to the emotional categories they believed each one of them belonged to. Each subject was presented with seven synthesized sentences for each one of the four training approaches (ED, EA, EI, EAT) and for each one of the four emotions (anger, fear, joy, sadness) (112 sentences in total).

Table 6: Emotion classification subjective evaluation of training HSMM approaches.

Emotional Categories	ED (%)	EA (%)	EI (%)	EAT (%)
Anger	97.3	97.6	97.9	98.5
Fear	66.7	67.3	69.3	72.0
Joy	81.8	83.3	84.8	86.3
Sadness	96.1	96.7	97.3	97.0

In Table 6 the results of the subjective test are shown. The experimental results show that the EAT models achieved the best emotion recognition rates followed by the EI and the EA models. The EAT models achieved 98.5%, 72%, 86.9% and 97% respectively for anger, fear, joy and sadness. Moreover the ED models presented the lowest performance in the emotion classification subjective evaluation test, achieving 97.3%, 66.7%, 81.8% and 96.1% respectively for anger, fear, joy and sadness. The reasoning behind these results lays on the basic principles of each method and the available training data. On the one hand, the ED models which are trained using the data of each emotional category, achieved the worst scores throughout all the emotional categories due to the inadequate amount of available training data. On the other hand, in the EA method, the available data of emotional speech are used only for adapting the neutral model to the target emotion. Consequently, the emotional speech data is adequate for the adaptation procedure, creating better models and managing to outperform the ED models. Moreover, in the EI method the average emotion model is build based on the clustering of the context-dependent emotion models (built in the training phase) to a shared decision tree, training more robust models, outperforming the EA models. Furthermore, in the EAT method, the MLLR adaptation which is used as an emotion normalization technique, managed to reduce the influence of emotion differences in respect to the EA method, building more robust models and achieving the best performance throughout all the emotional categories. Finally, it should be noted that anger and sadness emotional categories, as being more distinguishable (Table 3), managed to keep higher recognition rates in the subjective evaluation test throughout all the training approaches in contrast to joy and fear emotional categories in which cases the recognition rates dropped.

V. Conclusion

A comparison evaluation on HSMM training approaches in HMM-based speech synthesis for synthesizing emotional speech was presented in this paper. Since HMM do not support the explicit modelling of state durations in their training phase, the HSMM were used instead, in our experiments. A Modern Greek speech database of emotional speech, consisting of four categories of emotional speech: anger, fear, joy, and sadness, was used. Four training approaches of HSMM-based synthesis of emotional speech were presented and evaluated in this work.

A subjective classification test was performed evaluating the effectiveness of the training approaches to produce synthetic speech of specific emotions. The experimental results showed that the emotion adaptive training approach achieved the best emotion recognition rates followed by the emotion-independent and the emotion adaptation modelling approaches. Finally, the emotion-dependent modelling approach achieved the lowest performance.

References

[1] Hunt A, Black A. Unit selection in a concatenative speech synthesis system using a large speech database [C]. In: Proceedings of ICASSP, 1996, 373-376.

[2] Black A W, Cambpbell A W. Optimising selection of units from speech database for concatenative synthesis [C]. In Proceedings of EUROSPEECH, 1995, 581-584.

[3] Clark R A, Richmond K, King S. Multisyn: Opendomain unit selection for the Festival speech synthesis system [J]. Speech Communication, 2007, 49(4):317-330.

[4] Donovan R, Woodland P. A hidden Markov-modelbased trainable speech synthesizer [J]. Computer Speech and Language, 1999, 13(3): 223-241.

[5] Masuko T, Tokuda K, Kobayashi T, Imai S. Speech synthesis using HMMs with dynamic features [C]. In Proceedings of ICASSP, 1996, 389-392.

[6] Yoshimura T, Tokuda K, Masuko T, Kobayashi T, Kitamura T. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis [C]. In Proceedings of EUROSPEECH, 1999, 2347-2350.

[7] Yamagishi J, Nose T, Zen H, Ling Z H, Toda T, Tokuda K, King S, Renals S. A robust speaker-adaptive HMM-based text-to-speech synthesis [J]. IEEE Trans. Speech, Audio & Language Processing, 2009, 17(6):1208-1230.

[8] Zen H, Toda T, Nakamura M, Tokuda K. Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005 [J]. IEICE Trans. Inf. & Syst, 2007, E90-D(1):325-333.

[9] Zen H, Tokuda K, Black A W. Statistical parametric speech synthesis [J]. Speech Communication, 2009, 51(11):1039-1064.

[10] Fukada T, Tokuda K, Kobayashi T, Imai S. An adaptive algorithm for mel-cepstral analysis of speech [C]. In Proceedings of ICASSP, 1992, 137-140.

[11] Tamura M, Masuko T, Tokuda K, Kobayashi T. Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR [C]. In Proceedings of ICASSP, 2001, 805-808.

[12] Yamagishi J, Masuko T, Kobayashi T. MLLR adaptation for hidden semi-Markov model based speech synthesis [C]. In Proceedings of ICSLP, 2004, 1213-1216.

[13] Gauvain J, Lee C H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains [J]. IEEE Trans. Speech Audio Processing, 1994, 2(2):291-298.

[14] Ogata K, Tachibana M, Yamagishi J, Kobayashi T. Acoustic model training based on linear transformation and MAP modification for HSMM-based speech synthesis [C]. In Proceedings of ICSLP, 2006, 1328-1331.

[15] Nakano Y, Tachibana M, Yamagishi J, Kobayashi T. Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis [C]. In Proceedings of ICSLP, 2006, 2286-2289.

[16] Anastasakos T, McDonough J, Schwartz R, Makhoul J. A compact model for speaker-adaptive training [C]. In Proceedings of ICSLP, 1996, 1137-1140.

[17] Yamagishi J, Kobayashi T. Adaptive training for hidden semi-Markov model [C]. In Proceedings of ICASSP, 2005, 365-368.

[18] Tokuda K, Masuko T, Miyazaki N, Kobayashi T. Hidden markov models based on multi-space probability distribution for pitch pattern modeling [C]. In Proceedings of ICASSP, 1999, 229-232.

[19] Tokuda K, Yoshimura T, Masuko T, Kobayashi T, Kitamura T. Speech parameter generation algorithm for HMM-based speech synthesis [C]. In Proceedings of ICASSP, 2000, 1315-1318.

[20] Zen H, Tokuda K, Masuko T, Kobayashi T, Kitamura T. Hidden semi-Markov model based speech synthesis [C]. In Proceedings of ICSLP, 2004, 1180-1185.

[21] Yamagishi J, Onishi K, Masuko T, Kobayashi T. Modeling of various speaking styles and emotions

for HMM-based speech synthesis [C]. In Proceedings of EUROSPEECH, 2003, 2461-2464.

[22] Yamagishi J, Tamura M, Masuko T, Tokuda K, Kobayashi T. A context clustering technique for average voice model in HMM-based speech synthesis [C]. In Proceedings of ICSLP, 2002, 133-136.

[23] Yamagishi J, Kobayashi T. Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training [J]. IEICE Trans. Inf. & Syst, 2007, E90-D(2):533-543.

[24] Zervas P, Geourga I, Fakotakis N, Kokkinakis G. Greek Emotional Database: Construction and Linguistic Analysis [C]. In Proceedings of the 6th International Conference of Greek Linguistics, 2003.

[25] Lazaridis A, Kostoulas T, Ganchev T, Mporas I, Fakotakis N. VERGINA: A modern Greek speech database for speech synthesis [C]. In Proceddings of LREC, 2010, 117-121.

[26] Oatley K, Johnson-Laird P. The communicative theory of emotions [B]. In Human Emotions: A Reader, edited by J. Jenkins, et al. Oxford: Blackwell, 1998, 84-87.

[27] Mporas I, Ganchev T, Fakotakis N. A hybrid architecture for automatic segmentation of speech waveforms [C]. In Proceedings of ICASSP, 2008, 4457-4460.

[28] Wells J C. SAMPA computer readable phonetic alphabet [B]. In D., Gibbon, R., Moore, and R., Winski (eds.). Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter. Part IV, section B, 1997.

[29] Tokuda K, Zen H, Yamagishi J, Masuko T, Sako S, Black A, Nose T. The HMM-based speech synthesis system (HTS) Version 2.1, 2008, http://hts.sp.nitech.ac.jp/.

Authors' Profiles

Alexandros Lazaridis was born in Thessaloniki, Macedonia, Greece, in 1981. He graduated in September of 2005 from the Department of Electrical & Computer Engineering at Aristotle University of Thessaloniki, in Greece. He received his PhD at the Department of Electrical and Computer Engineering at the University of Patras, in February of 2011. Currently he is post-doctoral researcher at the University of Patras and non-tenured Lecturer at the University of Western Macedonia and non-tenured Assistant Professor at the Technological Educational Institute of Serres. He is author and co-author in more than 20 papers. His fields of research include Speech Processing, Voice Conversion, Speech Synthesis and Speech Prosody.

Iosif Mporas was born in Athens, Greece, in 1981. He graduated in 2004 (Diploma) from the Department of Electrical and Computer Engineering of the University of Patras, Greece. He received his PhD degree in July 2009. Currently he is post-doctoral researcher at the University of Patras and non-tenured Assistant Professor at the Technological Educational Institute of Patras. He is author and co-author in more than 50 publications in scientific journals and international conferences. His research interests include speech and audio signal processing, pattern recognition, automatic speech recognition, automatic speech segmentation and spoken language/dialect identification.

How to cite this paper: Alexandros Lazaridis, Iosif Mporas,"Evaluation of Hidden Semi-Markov Models Training Methods for Greek Emotional Text-to-Speech Synthesis", International Journal of Information Technology and Computer Science(IJITCS), vol.5, no.4, pp.23-29, 2013.DOI: 10.5815/ijitcs.2013.04.03