One-Class Classification for Spectral Join Cost Calculation in Unit Selection Speech Synthesis

Sotiris Karabetsos, *Student Member, IEEE,* Pirros Tsiakoulis, *Member, IEEE,*
Aimilios Chalamandaris, Spyros Raptis, *Member, IEEE*

***Abstract*—This letter introduces one-class classification as a framework for the spectral join cost calculation in unit selection speech synthesis. Instead of quantifying the spectral cost by a single distance measure, a data-driven approach is adopted which exploits the natural similarity of consecutive speech frames in the speech database. A pair of consecutive frames is jointly represented as a vector of spectral distance measures which provide training data for the one-class classifier. At synthesis runtime, speech units are selected based on the scores derived from the classifier. Experimental results provide evidence on the effectiveness of the proposed method which clearly outperforms the conventional approaches currently employed.**

***Index Terms*— Speech synthesis, Unit selection, One-Class classification, Join cost, Distance measures.**

I. Introduction

CORPUS-BASED unit selection, is currently the predominant approach for achieving high quality, near-natural speech synthesis. In principle, this method makes no explicit assumptions regarding the underlying speech model and relies on runtime selection and concatenation of speech units from a large speech database based on explicit matching criteria [1]-[3]. Among these criteria, the spectral join cost (spectral continuity cost) is considered as one of the most important. Its role is to determine how compatible candidate speech units are and whether they should be selected for concatenation or not during synthesis. The spectral join cost usually consists of a spectral similarity measure derived from parametric or non-parametric representations of the speech signal at the join boundaries [4]-[8]. Other measures are based on time-domain analysis in a transformed space, considering global boundary-centric optimization [9]. Although many spectral join cost measures have been proposed in the literature, there is still no widely acceptable solution to the optimal selection in terms of audible spectral discontinuities and human perception [7]. To cope with this, data-driven methods have been recently investigated. These methods include statistical modeling [8], [10], as well as approaches relying on human judgments [11], [12]. However, it is recognized that avoiding human judgment and following a machine learning framework which would solely rely on the available speech corpus (*speech database*), would be advantageous [1]. The speech database is fully comprised of naturally evolving adjacent speech frames, forming sequences of audibly "perfect" joins. On the other hand, "poor" joins are difficult and expensive to systematically collect and characterize. These conditions provide a strong motivation for adopting the One-Class Classification (OCC) framework which is well suited to address such problems [13], [14].

In this letter, a data-driven framework based on OCC is proposed and explored for the spectral cost estimation in unit selection speech synthesis. Instead of quantifying the spectral cost as a simple distance measure at the join boundary, we take advantage of the natural similarity of adjacent frames in the speech database. Hence, pairs of consecutive frames within specific speech units (e.g., per phoneme) are jointly represented as a feature vector comprising multiple distance measures between the two frames, which are then used to train the one-class classifier. During synthesis, a join between two candidate speech units is evaluated by forming a pair of speech frames at the concatenation boundaries, calculating the corresponding feature vector in a similar way and deriving a score from the classifier. Experimental results on vowel joins, using a unit selection Text-to-Speech (TTS) system for the Greek language, provide evidence on the effectiveness of the proposed method which clearly outperforms baseline approaches currently employed.

II. A Framework for Spectral Join Cost Computation Based on One-Class Classification

A. Overview

The spectral join cost is usually defined as a distance measure between spectral features representing the speech frames at the concatenation point. However, numerous studies employing several spectral representations and distance measures indicate that this approach is not optimal for predicting audible spectral discontinuities, thus highlighting the need for a more elaborate approach. To this end, the authors in [7] have shown that a weighted combination of different distance measures and speech representations performed better in detecting audible spectral discontinuities. In [12], a classifier is developed to detect audible

The authors are with the Voice and Sound Technology Department, Institute for Language and Speech Processing (ILSP) / Research Center "ATHENA", Artemidos 6 & Epidavrou, GR-15125, Athens, Greece, (e-mail: {sotoskar, ptsiak, achalam, spy@ilsp.gr}).

discontinuities based on subjective data derived from a listening experiment. However, the inherent difficulty in these methods is that they depend on data based on human assessments which are expensive and hard to collect. Hence, they are unable to exploit the "reference model" inherent in the available speech database. The machine learning framework offered by OCC is well suited for this purpose. OCC is employed in two-class pattern recognition problems where there is a plethora of data regarding the one class (*target-class*) whereas the other class (*non-target* or *outlier class*) is ill-defined. The scope of OCC is to act as a domain descriptor and to identify objects resembling the target class while rejecting all other cases [13], [14]. The problem of estimating the spectral join cost can be casted as a one-class problem, since perfect joins are readily available from the speech corpus which comprises "perfectly" joined, consecutive frames of natural speech. On the other hand, it is difficult and expensive to systematically collect and characterize a sufficient set of training examples of inappropriate joins with audible discontinuities. Furthermore, the problem of deriving an adequate feature set that would be able to efficiently represent successive speech frames and capture the underlying concatenation phenomenon while enhancing the separability of the target class, is far from trivial. Empirical and ad hoc approaches have often been employed for this purpose. In this work, we adopt a feature vector based on spectral distances as explained in the following subsection.

A schematic representation of the proposed OCC framework is given in Fig. 1. It relies on the pre-recorded and annotated speech corpus and consists of two stages, namely a *training stage* and an *evaluation stage*, both of which are performed offline. In the training stage shown in Fig. 1(a), phoneme-based speech signal extraction is performed. For each phoneme a parameterization step follows for the segmentation, analysis and joint representation of successive speech frames so as to construct the feature vectors for OCC training. In the evaluation stage, shown in Fig. 1(b), a similar procedure is followed, involving the frames to be concatenated among the candidate units. The resulting feature vector is fed to the classifier from which a decision is obtained. The decision may be either absolute (*hard*), namely a classification as *target* or *outlier* object, or in the form of a score (*soft*) usually expressed as a distance or a probability value (subsection II-C) [13].

B. Feature Extraction

The feature extraction procedure adopted in the present work is shown in Fig. 1(c). For every phoneme, each frame is analyzed into three spectral representations, namely Fourier power spectrum (*FFT)*, linear prediction coefficients (*LPC*) and Mel-frequency cepstral coefficients (*MFCC*). Then, every pair of adjacent frames is jointly represented as a feature vector consisting of five spectral distance measures computed from these representations. More specifically: (1) the *symmetrical Kullback-Leibler distance (KL-FFT)* on FFT-based power spectrum [5], [6], (2) the *symmetrical Kullback-Leibler distance (KL-LPC)* on LPC spectral envelopes [6], (3) the *Euclidean distance (E-MFCC)* on MFCC [5]-[8], (4) the *Mahalanobis distance* (*Mah-MFCC*) on MFCC [7], [8], and (5) the *Itakura distance* (*ITAK*) on LPC [15]. These representations together with the corresponding measures are widely employed in speech processing and their performance has been attested in several studies [15]. Each of these measures presents a number of advantages and drawbacks regarding their effectiveness in discriminating audible spectral discontinuities. Experimentation on using only a single distance measure to train the one-class classifier revealed no additional gain other than the one achieved by using that distance alone as a raw cost. Hence, to achieve enhanced performance by exploiting the OCC methodology, a more efficient feature vector should be utilized, thereby motivating the combination of multiple distance measures. Fusing these distinct measures into a single feature vector leads to a richer

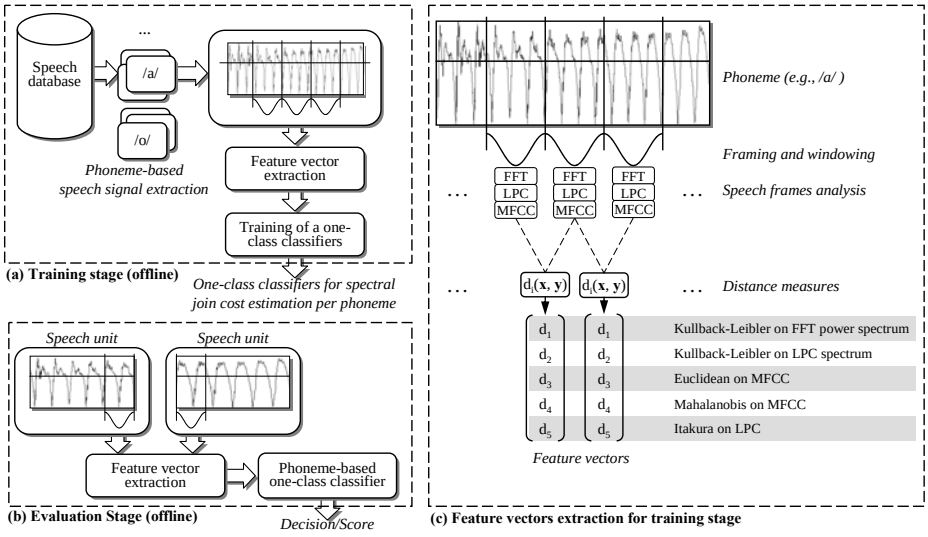

Fig.1. Overview of the One-Class classification framework for the spectral cost calculation in unit selection speech synthesis: (a) training stage, (b) evaluation stage and (c) feature extraction for joint representation of successive speech frames within a phoneme.

collective representation which combines their individual merits offering a more robust and complete "meta-measure". Moreover, an objective reference model of the *target class* is implicitly derived from the data, which is capable of discriminating whether the evolution of speech characteristics among two consecutive frames can be regarded as "natural" or not and if their deviation can be justified based on the underlying data. Such a data-driven approach is well motivated and emerges naturally in the context of unit selection speech synthesis.

C. One-Class Classifier

While several choices are available for the implementation of OCC, in this work we investigate the use of a *Gaussian Mixture Model One-Class Classifier* (*OCC-GMM*). The *OCC-GMM* is a density-based OCC, since the estimation of a probabilistic model for the *target class* is required. The *OCC-GMM* is derived as a linear combination of Gaussian probability density functions. If we let $\mathbf{x}$ represent an objects' feature vector, K be the number of Gaussians, $\boldsymbol{\mu}_k$ and Σ_k be the mean and the covariance matrix of the k-th Gaussian respectively, N the dimension of the data and a_k are the mixing coefficients, then the *OCC-GMM* is expressed as:

$$p_{OCC-GMM}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N}} \sum_{k=1}^{K} a_k \frac{1}{\sqrt{\det(\Sigma_k)}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right) \tag{1}$$

Classification is based on a threshold value θ_p, related to the proximity measure $p_{OCC-GMM}(\mathbf{x})$ which corresponds to a probability-based similarity measure of $\mathbf{x}$ to the *target class*. The threshold value is determined by specifying an acceptable error level on the training *target class* data. A new object $\mathbf{x}$ is labeled as a target object (*hard decision*), if the probability is larger than the threshold θ_p : $f(\mathbf{x}) = I(p_{OCC-GMM}(\mathbf{x}) > \theta_p)$ where I is an indicator function defined as: $I(A) = 1$, if A is true and 0, otherwise and $A = p_{OCC-GMM}(\mathbf{x}) > \theta_p$. More details can be found in [13], [14]. For the problem of detecting spectral discontinuities, a target object means that no audible discontinuity is present whereas an outlier object means the opposite situation. The quantity $p_{OCC-GMM}(\mathbf{x})$ can also be used directly as a score providing the means for a *soft decision*.

III. Experiments and Results

A. Experimental Setup

The speech database used for both the training and the evaluation experiments is part of a unit selection TTS system for the Greek language (*http://speech.ilsp.gr/TTS_samples/*). It is uttered by a female speaker and it consists of approximately one hour of a precisely annotated speech stimuli sampled at {16 KHz, 16 bits}. For the training stage, each instance of every phoneme is analyzed into 20ms frames with no overlap, followed by the computation of the quantities described in section II-B. Each frame is represented using an LPC order of 18, an FFT size of 512 and an MFCC vector of 12 coefficients excluding the zero one. The LPC spectral envelope is computed using a 512-point FFT. In all cases, a Hamming window is applied and a pre-emphasis filter using a factor of 0.97 is used. In this work, we restrict our analysis to vowels since spectral discontinuities are more evident in such phonemes [5]-[8]. A tolerable target rejection percentage of 15% was used to train the *OCC-GMMs* via a consistency-based model order selection criterion [16], resulting in a number of GMM components per phoneme in the order of 6-14, using the tools provided in [17]. In order to assess the proposed approach a procedure similar to the one described in [5], [6] and [12] was adopted. That is, a perceptual experiment was carried out in which five subjects with background in speech synthesis were asked to express their opinion on a binary decision task regarding the detection of audible spectral discontinuity in vowel concatenations. The test stimuli was comprised of 964 C_iVC_j (*C: consonant set*, *V: vowel set*) utterances and it was created by concatenating diphones of the form C_iV and VC_j from the speech database, ensuring a vowel duration of at least 150 msec as well as small intensity and pitch deviations. The experiment was divided into three blocks performed in three hourly sessions. A familiarization phase preceded each block. The participants were instructed to use headphones and to focus only on vowel transitions. A concatenation was characterized either as target (*continuous*) or outlier (*discontinuous*) depending on the majority of the subjects' decisions. The results of the test were utilized for the performance evaluation of the one-class classifier. In order to relate the listening test results not only with the *OCC-GMM* but also with each separate spectral distance measure employed, a procedure similar to the one described in [5], [6] and [12] was followed, based on receiver operating characteristic (ROC) curves. For each measure D, the probability density functions, $p(D|target)$ and $p(D|outlier)$ are estimated. Based on the test results, the first one denotes the probability of a measure given that a join was characterized as target, while the latter denotes the probability of a measure given that the join was characterized as outlier. Hence, based on the problem definition and for a specific threshold value α, the *true positive fraction* (TP), that is the case where a join is correctly classified as target, against the *false positive fraction* (FP), that is the case where a join is incorrectly classified as target, are expressed as:

$$TP(\alpha) = \int_{-\infty}^{\alpha} p(D|target)\, dD \tag{2}$$

$$FP(\alpha) = \int_{-\infty}^{\alpha} p(D|outlier)\, dD \tag{3}$$

In addition, the case of a join incorrectly classified as outlier (*false negative fraction*) is *FN=1-TP* and the case of a join is correctly classified as outlier (*true negative fraction*) is *TN*=1-*FP*. For a varying threshold α, a ROC curve is formed by plotting *TP* against *FP*.

B. Results and Discussion

Fig. 2 depicts the overall phoneme-independent ROC curves

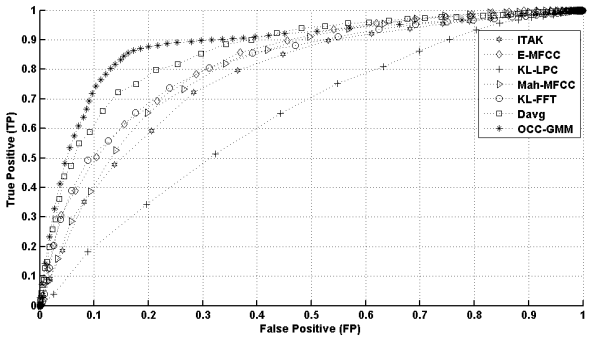

Fig. 2. Comparative ROC curves for the performance evaluation of the *OCC-GMM* approach.

of the *OCC-GMM* compared to the ones regarding each distance measure alone, as well as their equally weighted (average) combination denoted as D_{avg}. It is observed that the proposed approach provides a significant performance improvement over each separate measure. For an FP of 10% a TP of approximately 70% is achieved denoting an improvement in the order of 20% over KL-FFT and E-MFCC which are the two best performing measures. Moreover, these results indicate a 90% correct detection of spectral discontinuities over about a 30% penalty of missing a target object. In addition, by exploiting *OCC-GMM* an improvement in the order of 10% is achieved over D_{avg}.

TABLE I
OCC-GMM COMPARATIVE PERFORMANCE RESULTS PER PHONEME IN TERMS OF ACHIEVED TP FOR AN FP OF 10%

	Phoneme				
Measure	/a/	/e/	/o/	/i/	/u/
OCC-GMM	**74%**	**70%**	**59%**	**84%**	**86%**
E-MFCC	54%	55%	34%	58%	34%
KL-FFT	53%	51%	43%	42%	59%
ITAK	54%	30%	25%	62%	26%
Mah-MFCC	49%	52%	23%	51%	33%
KL-LPC	26%	15%	23%	36%	10%

Table I provides the *OCC-GMM* performance results per phoneme in terms of achieved TP for an FP of 10%, together with the corresponding results obtained from every individual spectral distance measure alone. In all cases, it is seen that the *OCC-GMM* performs significantly better than each measure both in identifying proper joins and in detecting audible spectral discontinuities. Moreover, it is observed that no single measure performs well in this task, a result which is consistent with previous studies [5]-[8].

TABLE II
OCC-GMM TP SCORES PER PHONEME FOR DIFFERENT SPEECH CORPUS SIZE (FP=10%)

	Phoneme				
Corpus Size	/a/	/e/	/o/	/i/	/u/
100%	74%	70%	59%	84%	86%
50%	73%	65%	57%	81%	82%
20%	70%	63%	55%	79%	65%

In addition, the effect of the training corpus size on the performance of the OCC-GMM is shown in Table II. Given an FP of 10% the TP is recorded when different sizes of the training corpus are considered. It is seen that as the size of the training corpus increases, enhanced performance is obtained.

IV. CONCLUSIONS AND FURTHER WORK

In this work, one-class classification is proposed as a new paradigm for the spectral cost calculation in unit selection speech synthesis. The essence of exploiting OCC in this context is that it provides an effective data-driven mechanism that overcomes the inherent shortcomings of conventional approaches. This is confirmed by the experimental results where a significant improvement over baseline techniques was recorded in detecting audible spectral discontinuities. Ongoing work aims to further elaborate this framework, also in terms of finding an optimal feature set, as well as to investigate whether OCC can be employed for the computation of the total concatenation cost.

REFERENCES

[1] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.
[2] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *in IEEE Int. Conf. Acoustics, Speech and Signal Processing-ICASSP*, 1996, pp. 373–376.
[3] J. Wouters and M. Macon, "Control of spectral dynamics in Concatenative speech synthesis," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 1, pp. 30–38, Jan, 2001.
[4] J. Wouters and M. Macon, "A perceptual evaluation of distance measures for concatenative speech synthesis," *in Proc. ICSLP'98*, vol. 6, Sydney, Australia, 1998, pp. 2747–2750.
[5] Y. Stylianou and A. K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," *in IEEE Int. Conf. Acoustics, Speech and Signal Processing-ICASSP*, 2001, pp. 837–840.
[6] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 1, pp. 39–51, Jan. 2001.
[7] J. Vepa and S. King, "Join cost for unit selection speech synthesis," in *Text to Speech Synthesis: New Paradigms and Advances*, S. Narayanan and A. Alwan, Eds. NJ: Prentice-Hall, pp. 35–62, 2004.
[8] J. Vepa and S. King, "Subjective Evaluation of Join Cost and Smoothing Methods for Unit Selection Speech Synthesis," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1763-1771, Sep. 2006.
[9] J. R. Bellegarda, "A Global Boundary-Centric Framework for Unit Selection Text-to-Speech Synthesis," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 990-997, May 2006.
[10] P. Taylor, "Unifying unit selection and hidden Markov model speech synthesis", *in Proc. of Interspeech 2006*, Pittsburgh, USA, Sept. 2006.
[11] A. K. Syrdal and A. D. Conkie, "Data-Driven Perceptually based Join Costs," *in Proc. 5th ISCA Speech Synthesis Workshop*, Pittsburgh, June 2004, pp. 49-54.
[12] Y. Pantazis, Y. Stylianou, and E. Klabbers, "Discontinuity detection in concatenated speech synthesis based on nonlinear speech analysis," *in Proc. of Interspeech 2005*, Lisbon, Portugal, 2005, pp. 2817-2820.
[13] D. M. J. Tax, *One-class classification; Concept-learning in the absence of counter-examples*. Ph.D. thesis, ISBN: 90-75691-05-x, Delft University of Technology, 2001.
[14] M. Markou and S. Singh, "Novelty detection: A review-part 1: Statistical approaches," *Signal Processing*, vol. 83, no. 12, pp. 2481–2497, 2003.
[15] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ, 1993.
[16] D. M. J. Tax and K. R. Mueller, A Consistency-Based Model Selection for One-class Classification, *in Proc. ICPR 2004*, vol. 2, Cambridge UK, IEEE Computer Society, 22-26 August, 2004, pp. 363-366.
[17] D. M. J. Tax, *DDtools, the Data Description Toolbox for Matlab*, version 1.7.3, 2009.