ISSN 0005 -1144
ATKAAF 52 (2), 76-184 (2011)
UDK: 007:62/68
Godina 52, Broj 2, Str. 76- 184
Zagreb, Travanj - Lipanj 2011

• automatika • mjerenje • elektronika • računarstvo • komunikacije • automatika • m

• measuring • electronics • computing • communications • automation • measuring

• elektronika • računarstvo • komunikacije • automatika • mjerenje • elektronika • ra

• computing • communications • automation • measuring • electronics • computing

• komunikacije • automatika • mjerenje • elektronika • računarstvo • komunikacije •

časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije
journal for control, measurement, electronics, computing and communications

Izdaje / Published by KoREMA **Zagreb, Hrvatska / Croatia**

52 automatika

GODINA/VOLUME 2/2011

AUTOMATIKA

Journal for control, measurement, electronics, computing and communications
ISSN 0005-1444, UDK 007:62/68
ATKAAF 52(2), 76–184 (2011)
Volume 52, No. 2, April-June 2011

Published quarterly by KoREMA – Croatian Society for Control, Communications, Computing, Electronics, Measurement and Control, Member of IFAC and IMECO

Partly financed by the Ministry of Science, Education and Sports of the Republic of Croatia

Abstracted/Indexed in Science Citation Index Expanded, EBSCO and INSPEC (Institution of Electrical Engineers)

EDITORIAL OFFICE AND JOURNAL ADMINISTRATION

Journal Preparation for Print: **Ivan Marković**, coordinator, **Marija Đakulović**, **Tamara Hadjina**, **Vlaho Petrović**
Print: **Europrint**, 10410 Velika Gorica, Kneza Porina 28, Croatia
Manuscript Submission System: **Srećko Jurić-Kavelj**
Journal subscription and Advertisement announcement information: **Marina Štimec**, KoREMA Secretary

KoREMA Secretariat and Journal Address:
KoREMA (for journal Automatika), Unska 3, 10410 Zagreb, Croatia
Phone: ++(385 1) 6129 869, Fax: ++(385 1) 6129 870
E-mail: korema@korema.hr, URL: http://automatika.korema.hr

automatika

JOURNAL FOR CONTROL, MEASUREMENT, ELECTRONICS, COMPUTING AND COMMUNICATIONS

Published by KoREMA
Member of IFAC and IMEKO

Zagreb, Croatia

ISSN 0005–1144
ATKAAF 52(2), 76–184 (2011)
UDK: 007:62/68

Volume 52–2011
No 2

CONTENTS

ISSN 0005-1144
ATKAFF 52(2), 132–146(2011)

Antonio Vasilijević, Davor Petrinović

Perceptual Significance of Cepstral Distortion Measures in Digital Speech Processing

UDK 004.934'1
IFAC 5.9.3

Original scientific paper

Currently, one of the most widely used distance measures in speech and speaker recognition is the Euclidean distance between mel frequency cepstral coefficients (MFCC). MFCCs are based on filter bank algorithm whose filters are equally spaced on a perceptually motivated mel frequency scale. The value of mel cepstral vector, as well as the properties of the corresponding cepstral distance, are determined by several parameters used in mel cepstral analysis. The aim of this work is to examine compatibility of MFCC measure with human perception for different values of parameters in the analysis. By analysing mel filter bank parameters it is found that filter bank with 24 bands, 220 mels bandwidth and band overlap coefficient equal and higher than one gives optimal spectral distortion (SD) distance measures. For this kind of mel filter bank, the difference between vowels can be recognised for full-length mel cepstral SD RMS measure higher than 0.4 - 0.5 dB. Further on, we will show that usage of truncated mel cepstral vector (12 coefficients) is justified for speech recognition, but may be arguable for speaker recognition. We also analysed the impact of aliasing in cepstral domain on cepstral distortion measures. The results showed high correlation of SD distances calculated from aperiodic and periodic mel cepstrum, leading to the conclusion that the impact of aliasing is generally minor. There are rare exceptions where aliasing is present, and these were also analysed.

Key words: Aliasing, Digital speech processing, MFCC, Mel cepstrum, SD Measure, Speech recognition

Percepcijska utemeljenost kepstranih mjera udaljenosti za primjene u obradi govora. Jedna od danas najčešće korištenih mjera u automatskom prepoznavanju govora i govornika je mjera euklidske udaljenosti MFCC vektora. Algoritam za izračunavanje mel frekvencijskih kepstralnih koeficijenata zasniva se na filtarskom slogu kod kojeg su pojasi ekvidistantno raspoređeni na percepcijski motiviranoj mel skali. Na vrijednost mel kepstralnog vektora, a samim time i na svojstva kepstralne mjere udaljenosti glasova, utječe veći broj parametara sustava za kepstralnu analizu. Tema ovog rada je ispitati usklađenost MFCC mjere sa stvarnim percepcijskim razlikama za različite vrijednosti parametara analize. Analizom parametara mel filtarskog sloga utvrdili smo da filtar sa 24 pojasa, širine 220 mel-a i faktorom preklapanja filtra većim ili jednakim jedan, daje optimalne SD mjere koje se najbolje slažu s percepcijom. Za takav mel filtarski slog granica čujnosti razlike između glasova je 0.4-0.5 dB, mjereno SD RMS razlikom potpunih mel kepstralnih vektora. Također, pokazat ćemo da je korištenje mel kepstralnog vektora odrezanog na konačnu dužinu (12 koeficijenata) opravdano za prepoznavanje govora, ali da bi moglo biti upitno u primjenama prepoznavanja govornika. Analizirali smo i utjecaj preklapanja spektara u kepstralnoj domeni na mjere udaljenosti glasova. Utvrđena je izrazita koreliranost SD razlika izračunatih iz aperiodskog i periodičkog mel kepstra iz čega zaključujemo da je utjecaj preklapanja spektara generalno zanemariv. Postoje rijetke iznimke kod kojih je utjecaj preklapanja spektara prisutan, te su one posebno analizirane.

Ključne riječi: preklapanje spektara, digitalna obrada govora, MFCC, mel kepstar, SD mjera, prepoznavanje govora

1 INTRODUCTION

Until the 1990s, the common method used to measure the quality of speech was by conducting subjective tests [1], [2]. Design of the objective measures that correlate well with subjective results is of prime importance in order to eliminate expensive and time consuming listening tests. The most common objective measures used for assessing the subjective quality of speech are based on perceptual auditory models. These objective measures quantify the perceived quality of distorted speech relative to an undistorted reference sample. Some of the known methods for objective evaluation of speech quality are Bark spectral density (BSD) [2], perceptual speech quality measure (PSQM), measuring normalizing blocks (MNB), perceptual analysis measurement system (PAMS), and perceptual evaluation of speech quality (PESQ) [3], [4]. These methods are not only used for evaluation of speech quality but also for evaluation of audio quality and quality of service (QoS) in VoIP [5]. Classical objective quality estimators based on SNR do not, in general, provide useful estimates of the perceived speech quality. Methods like cepstral distance and Bark spectral distortion are much better estimators but these methods still lack reliability when the broadest class of conditions is considered. The PSQM and the MNB estimators provide good results in many different applications [6], [7].

The objective measure is supposed to accurately represent perceptual similarity between the speech segment and the reference model. Quality of the measure depends on the feature vector used. Typically, some aspects of the feature vector, representing the speech spectral envelope, should be emphasized and the less relevant aspects ignored. A good measure would emphasize similarities in spectral peak positions and deemphasize the higher frequency content while ignoring spectral tilt and low-amplitude components.

Early automatic speech recognition (ASR) algorithms used simple filter banks or discrete Fourier's transformation (DFTs) to obtain the needed features. From the early 1970s to the mid-1980s, linear predictive coding (LPC) coefficients and their transformations (e.g. reflection coefficients, Line Spectral Frequencies LSFs) were widely used in ASR as well. Since the mid-1980s, the most widely used feature vector for ASR has been the MFCC [8]. The original MFCC algorithm was introduced by Davis and Mermelstein in 1980 [9]. It combines perceptually spaced filter bank with the discrete cosine transformation (DCT). Resulting feature vector of a dozen parameters is a compact representation of the voice segment. Based on the MFCC feature vector, objective measures of sound distances were developed.

Even though the mel cepstral distortion measure is one of the most commonly used measures in digital speech processing (ASR [10], [8], speaker recognition [11], [12] speech reconstruction [13], [14], speech synthesis, text-to-speech (TTS) [8]), it still represents a rough estimation of perceptual distinction between two phonemes.

This work analyses the relation between different mel cepstral representations of speech utterance and human perception of phonemes' distances. There are a number of parameters used in mel cepstral analysis which can affect the mel cepstral measure. First we analysed the influence of mel filter bank parameters, such as number of channels, bandwidth and band overlap factor, on the mel cepstral measure. The bandwidth was varied from 120 mels to 410 mels, number of filter channels from 12 to 46 bands and overlap factor from 2/3 to 8.

The next point in our research was to try to justify usage of the truncated mel cepstral vector (first 12 coefficients) instead of the full-length vector as a feature vector to calculate objective measure. Measures computed from both types of vectors were compared with subjective perceptual results.

The third analysed aspect was the impact of aliasing in cepstral domain on cepstral distortion measures. For that purpose "aperiodic" and "periodic" mel cepstral vectors were created. Aperiodic (infinite) mel cepstral vector was computed from the continuous mel-spectrum that is a continuous function of the filter band central frequency (i.e. infinite number of filter channels). Periodic mel cepstral vector is obtained by spectral sampling, i.e. by periodic expansion of the aperiodic vector with the period determined by the finite number of filter channels. Such aliasing introduces ambiguity into cepstral representation of speech, since perceptually different speech sounds might produce identical cepstral vectors. Section 2 reviews the methods used in this research with all steps involved. Section 3 presents the algorithm used to calculate all feature vectors and corresponding objective measures. Section 4 discusses the results for all three experiments, while section 5 presents the conclusion.

2 METHODOLOGY

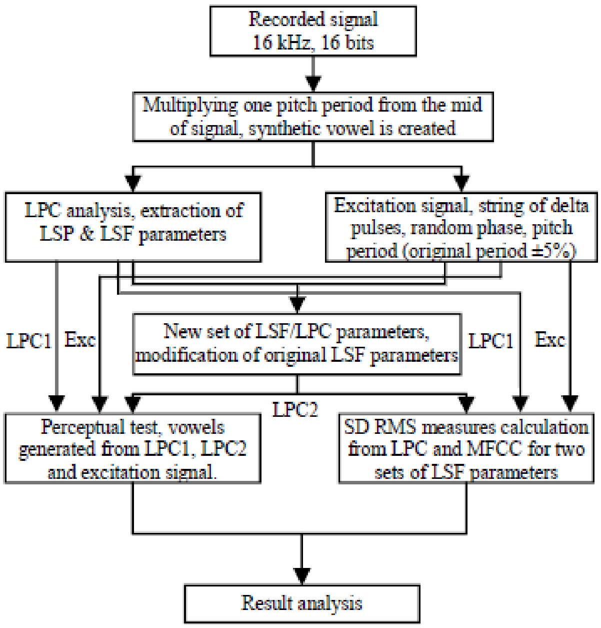
Recorded signal
16 kHz, 16 bits
Multiplying one pitch period from the mid of signal, synthetic vowel is created
LPC analysis, extraction of LSP & LSF parameters
Excitation signal, string of delta pulses, random phase, pitch period (original period ±5%)
LPC1
Exc
New set of LSF/LPC parameters, modification of original LSF parameters
LPC1
Exc
LPC2
Perceptual test, vowels generated from LPC1, LPC2 and excitation signal.
SD RMS measures calculation from LPC and MFCC for two sets of LSF parameters
Result analysis

Fig. 1. Methodology

2.1 Data base

Subjective testing methodology used in our work required synthetic phoneme with controlled variation of the spectral envelope, which can only be achieved by using vowels. Controlled variation of the shorter phoneme's spectral envelope would be very difficult due to fast variation and shorter stationary period of it's spectral envelope. For this reason and in order to perform a perceptual validation of distortion measures we have created a database limited to a synthetic isolated Croatian vowels. For each vowel, one reference stimulus was synthesized together with a large number of modified stimuli. Technique that was used for the syntesis of the reference and modified stimuli was identical. The only difference between them was the spectral envelope. For the reference vowel the envelope was estimated from the recorded speech utterance of that particular vowel to achieve naturalness. Spectral envelope estimation was based on the conventional LPC analysis, giving the time-invariant all-pole model of that particular vowel. Estimated model was then manipulated in order to create a modified stimuli. The goal was to create a set of modified spectral envelopes which are all close to the threshold of just perceivable difference from the referent one. Reference and modified envelopes were then used to create synthetic vowels, by exciting the corresponding LPC filters with identical quasi-periodic excitation. Excitation was formed as a band-limited pulse train generated according to the sinusoidal model. In order to achieve naturalness of synthetic vowels, pitch contour of the excitation signal was varied linearly in time from -5% to $+5\%$ of the estimated pitch period of extracted speech segment. Introduced pitch variations are also helpful to allow for different interactions between the LPC model (formant structure) and the fine (pitch) spectral structure of the synthetic vowel. Such interactions are particularly important since they affects both objective and perceptual differences. Slight frequency dependent phase dispersion of sinusoidal components was also introduced to improve the naturalness of synthesized vowels (equation (4)).

2.2 Perceptual experiment

Created database was used to setup a simple perceptual experiment. Chosen test type was a modified version of three-alternative forced choice test (3AFC). The subject was presented with a sequence of three stimuli, i.e. three synthesized vowels with short pauses in between. The subject was told that two vowels are always identical and that he/she has to choose the different one (A, B or C). Pair of stimuli used in each test utterance was formed from two synthetic vowels. One was always the reference vowel (R), while other was randomly chosen from the database of modified vowels (M). Presentation order was randomized with possible sequences being MRR, RMR or RRM. Unknown to the subject, the "all-reference" sequence RRR was intermittently included in the test as a hidden anchor, to check for consistency or possible bias. Modification to the conventional 3AFC test made in our experiment was the inclusion of one additional answer, "NONE". Subjects were instructed to use this answer only if the different vowel couldn't be identified by no other means except by guessing. Such answers were counted as three trials with one correct and two wrong answers, what corresponds to the guessing probability. Subjects were also allowed to repeat the same utterance several times before making the final guess.

2.3 Objective distance evaluation

For each vowel pair used in perceptual tests, several objective distance measures were computed. These include the simple Spectral Distance (SD) between log-magnitude responses of the referent and modified LPC filter and several cepstrum based distance measures computed from synthesized vowels for different parameter values of these measures (e.g. number of analysis channels and their bandwidth, cepstrum truncation, etc.). Synthetic nature of utterances used in our setup allows for precise control of objective distances. The referent LPC model can be manipulated in a way to obtain a pair of vowels on a certain, desired objective distance, measured by a chosen distance metric, i.e. MFCC SD measure. It would have been ideal to modify the MFCC vector directly and then generate the second vowel from the modified MFCC vector. However, two steps of the MFCC algorithm are irreversible which makes such direct approach impossible. These are computation of the magnitude of the complex speech spectrum which discards the phase information and mel filtering (smoothing) that nonlinearly reduces the frequency resolution of the speech power spectrum. To circumvent this problem, we have used a model based approach by means of the Line Spectral Frequencies (LSF) to generate the modified synthetic vowel. LSF parameters can be easily calculated from the referent set of LPC parameters. Genetic algorithm or manual manipulation were used to change them and create appropriate sets of new LSF parameters for modified vowels with desired objective distances from the referent one. The basic idea used in our experiment is inherited from the speech coding domain, where similar objective and subjective measures are used to evaluate the quantization accuracy of the speech spectral envelope. The LSF parameters are one of the most commonly used LPC representations in speech coding. Thus, the effect of quantization induced LSF variation on the subjective quality of coded speech is extensively studied in the speech coding literature. Commonly used limen for "transparent" quantization for quantizer design is the average SD distance of 1dB. However, this rather simple objective metrics does

Table 1. Vowels distinction levels

Score	Symbol	Level of distinction	Correctly guessed vowel
1.	Circle	Imperceptible	0% -50%
2.	Square	Barely perceptible	50% - 90%
3.	Triangle	Perceptible	90% - 100%

not correlate well with the human perception. Therefore, more elaborate distance measures that mimic psychological process of hearing are used in speech recognition tasks (e.g. MFCC SD measures).

2.4 Analysis of perceptual experiments

Our main goal was to compare the results of the perceptual experiment to the commonly used objective distance measures and to fine tune the parameters of MFCC based measure to achieve the best compliance with perceptual results. Ideally, such comparison should be performed by estimating the psychometric function that models the probability of correct identification of modified vowel as a function of a chosen objective distance that measures the perturbation level introduced into the spectral envelope of a modified vowel. Due to the "many-to-one" nature of all objective measures, infinite number of different envelope perturbations give identical objective distances, although the perceptual responses might be quite different. Therefore, the perceptual compliance of the chosen objective measure can be found as a correlation between the objective distance and measured perceptual response. For the good objective measure the perceptual results (i.e. measured probability of correct guesses) must lie close to a smooth sigmoid curve, that models the ideal psychometric response. In order to obtain sufficient reliability of perceptual identification probability many trials are necessary, with large number of test subject. In our limited experiment, we couldn't afford such exhaustive perceptual testing. Therefore, perceptual responses of vowel distinction were classified into only three classes, based on 30 trails per each modified vowel. These classes are: Imperceptible, Barely Perceptible and Perceptible with corresponding probabilities given in table 1.

2.5 Comparison to the objective measures

We were particularly interested in identifying and studying modified vowels with strong disagreement between different objective measures, i.e. when one measure claims that the modification must be perceivable, while according to the other, the distance to the referent vowel is supposed to be small. Perceptual tests for such ambiguous vowels give the true answer. Special attention was given to such unusual modifications in order to investigate a type of perceptually relevant modifications that might get unnoticed in the commonly used distance measures.

Results of our experiments are mostly presented as two-dimensional graphs in order to make pairwise comparisons of different objective measures. Perceptual responses for individual modified vowels are shown as data point in these plots, with three symbols corresponding to the three described classes (table 1). Position of each point is given by the values of two chosen objective measures. Grouping of data points belonging to the same perceptual class along one or the other axis clearly demonstrates the perceptual compliance of a chosen measure.

2.6 Parametric representation for synthetic vowels

Database of synthetic vowels was created from the two characteristic male Croatian vowels "A" and "U". These are similar to English vowels in *sun* and *soon*. Speech was sampled with 16 kHz sampling frequency and 16 bits amplitude resolution. We have applied basic band-pass filtering to the signal to remove DC component and high-frequency regions affected by aliasing. In order to find parametric representation of the vowel, we have used conventional LPC analysis [10] performed on a central segment of the recorded vowel. In early ASR algorithms the LPC coefficients α_i directly served as ASR feature vectors [15]. LPC parameters capture characteristics of the vocal tract and accurately describe the envelope of the speech power spectrum. In our experiment the LPC prediction order was set to P=28 in order to obtain a precise spectral model of the wide-band speech signal (300 Hz-7.5 kHz). The optimal predictor was determined using the covariance method [10] and was additionally checked for stability and minimum formant bandwidth. Obtained LPC predictor coefficients α_i are not suitable for direct filter modification. They represent coefficients of the Direct-form IIR filter implementation and exhibit very complicated spectral sensitivity behaviour. A small change of a single coefficient typically changes the whole transfer function and can even result in an unstable filter. Therefore, the LPC model was transformed to the equivalent representation based on the Line Spectral Pairs (LSP) or Line Spectral Frequencies (LSF) that were first introduced by Itakura [16] and are broadly used in speech coding. The LSP coefficients are also effective features to discriminate between speakers and can be used in speaker recognition applications for person identification and verification [17]. The LSF parameters can be found from the LPC parameters α_j as the complex roots $Z_j = e^{i\omega_j}$ of two polynomials:

$$Q(z) = 1-(\alpha_1-\alpha_p)z^{-1}-\cdots-(\alpha_p-\alpha_1)z^{-p}-z^{-(p+1)} \quad (1)$$

$$P(z) = 1-(\alpha_1+\alpha_p)z^{-1}-\cdots-(\alpha_p+\alpha_1)z^{-p}+z^{-(p+1)} \tag{2}$$

In order to ensure stability of the LPC filter, roots of $P(z)$ and $Q(z)$ must have the following properties: all roots are distinct and lie on the unit circle and the roots of $P(z)$ are interlaced with those of $Q(z)$. Therefore, if LSF parameters of the stable LPC filter are sorted in ascending order, then odd frequencies correspond to one of the polynomials, while even frequencies correspond to the other:

$$0 < \omega_1 < \omega_2 < \cdots < \omega_{p-1} < \omega_p < \pi \tag{3}$$

While the values of the individual LPC parameters do not tell us much about spectral envelope and formants, LSF parameters clearly characterize the central frequency and bandwidth of a given formant. Change of a given LSF produces a change in the LPC power spectrum only in the neighbourhood of that particular LSF. The property of localized spectral sensitivity of LSF parameters makes them an excellent choice for spectral envelope manipulation. Additionally, stability of the modified filter can be easily ensured by simply sorting the LSFs after modification.

2.7 Vowel modification with genetic algortihm

By modifying LSF parameters of the referent vowel, we can generate new vowel with desired MFCC feature vector. If only a local spectral modification is required, e.g. one formant change, then LSF parameters defining that particular formant can be tuned manually. If we want to generate vowel which is on a certain distance (measured by SD mel cepstral measure) from the reference vowel, then simple Genetic algorithm (GA) can be applied. Genetic algorithm uses natural evolution mechanisms to generate optimum or close to optimum result.

In our experiment, we have used GA to obtain desired MFCC feature vectors by modifying the referent LSF parameters. A set of LSF parameters represents a string which is processed by GA. Starting population consists of ten strings. Next generation of strings is usually obtained through sequence of reproduction, crossover and mutation but crossover was not used in our research. Reproduction is responsible for survival of the fittest and removing of the poorest. The ‘goodness’ of a particular string is evaluated by the closeness to the desired MFCC feature vector. Three fittest strings are reproduced to the next generation without changes and the rest of strings are removed. The seven new strings of the next generation are generated by mutating the three fittest strings of a current generation. The algorithm is stopped when the set of LSF parameters with required feature vector is found (or at least, close enough). If desired set of parameters is not found, the algorithm is stopped after a fixed number of iterations.

2.8 Vowel synthesis

From these two sets of LPC parameters, the referent one and the one obtained by GA modification of referent parameters, two vowels are generated. The duration of the synthesized vowels is fixed at 300 ms. The common signal used to excite two all-pole LPC filters is defined as follows:

$$u(n) = \sum_{k=0}^{K} \frac{1}{K+1} cos(\omega_0 kn(1+\Delta f - \frac{2\Delta f n}{N-1}) + \phi_k)$$
$$0 \le n \le N-1 \tag{4}$$

$\omega_0 = 2\pi/pp$ is the fundamental frequency, while pp is the corresponding referent pitch period expressed in samples. Random frequency weighted phase dispersion is added to each of the harmonics, $\phi_k = (\pi/2)\cdot rand \cdot (k/K)$ where $rand$ is a random number $-1 \le rand \le 1$ defining the initial phase of the k^{th} harmonic. Fundamental frequency is changed linearly throughout the duration of the vowel as: $\omega_0 \cdot (1+\Delta f - 2\cdot\Delta f \cdot n/(N-1))$. The parameter $\Delta f = 0.05$ is chosen such that fundamental frequency deviation at vowel boundaries is $\omega = \omega_0 \pm 5\%$. After synthesis a short fade-in and fade-out intervals are imposed on the vowels, to avoid any audible clicks at the beginning or at the end that could be used by the listeners as unwanted perceptual cues to discern between the two vowels.

2.9 Results analysis, clustering and visualizations

Results of the described perceptual experiment were used as a reference data set for objective distance measures. Objective measures depend on the analysis parameters and feature vector choice. The aim of our investigation was pairwise comparison of different objective measures relative to the subjective results. Therefore, the distance between each referent and modified vowel pair used in subjective evaluation, was computed using two selected objective measures and the results were plotted in the plane formed by these two objective measures. For a good objective measure, the data points of the three subjective score classes (Table 1) must be clearly grouped along one (or both) axes in this plane. In order to easily identify such groupings the K Nearest Neighbours (KNN– 3) classification [18] was used to divide the distance plane in a way that preferably only samples from one class are in each separated region. Our aim was to classify the areas in these plots where pairs of vowels are distinguishable or indistinguishable through listening tests. Every vowel pair used in the perceptual experiment represents one sample of the KNN training set. Information about classes (perceptual test scores) and x-y coordinates (two different SD measures) of the training set members are available and used for training. The KNN– 3 classifier classifies the new sample into the majority class of the three nearest neighbours from the training set, measured by Euclidean distance. In the case that all three nearest neighbours are from different classes, the class of the nearest neighbour will be assigned to the new sample. Finally, the borders of identified clusters are superimposed onto the distance graph to facilitate visual inspection of the sample grouping. Shape of the border between individual classes tells us about the quality of certain objective measure. Straight horizontal or vertical borders show that particular objective measure is fully compatible with subjective measure.

3 EXTRACTION OF THE FEATURE VECTORS AND OBJECTIVES MEASURES

The two vowels: referent and modified have different spectral envelopes that are modelled with corresponding LPC filters. The simplest distance measure between the two vowels is the conventional SD distance that is often used for evaluation of the speech coding performance, related to the quantization of the LPC filter parameters. It is equal to the simple RMS value of the log-magnitude difference between frequency responses of the two LPC filters above the linear frequency axis.

$$\mathrm{SD_{LPC}} = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K-1}(20(\log_{10}|H_1(k)|-\log_{10}|H_2(k)|))^2} \tag{5}$$

Where $H(k)$ is a sampled magnitude response of the all-pole LPC filter that models the speech spectral envelope. Since the LPC parameters are held constant for the whole duration of vowels, the distance between the referent and modified vowel is described with a single SD value. Remaining mel-cepstral based objective measures investigated in this paper are derived from the actual waveforms of synthetic vowels, by means of short time analysis as it is usually done in typical applications. Therefore, a sequence of distance values is obtained, one for each analysis frame, that are averaged across frames resulting with the average distance:

$$\overline{\mathrm{SD}}_{\mathrm{cepstrum}} = \frac{1}{M}\sum_{m=1}^{M}\mathrm{SD}_{\mathrm{cepstrum}} \tag{6}$$

where M is the number of analysis frames within one vowel. Averaging helps to reduce for possible uncertainties due to interaction of the spectral envelope and excitation signal.

3.1 MFCC analysis algorithm

Although the MFCC analysis and corresponding distance measures are well known in the ASR literature [10] in this section we will revisit the main expression and introduce crucial parameters that affect compliance of such objective measures with perceptual results.

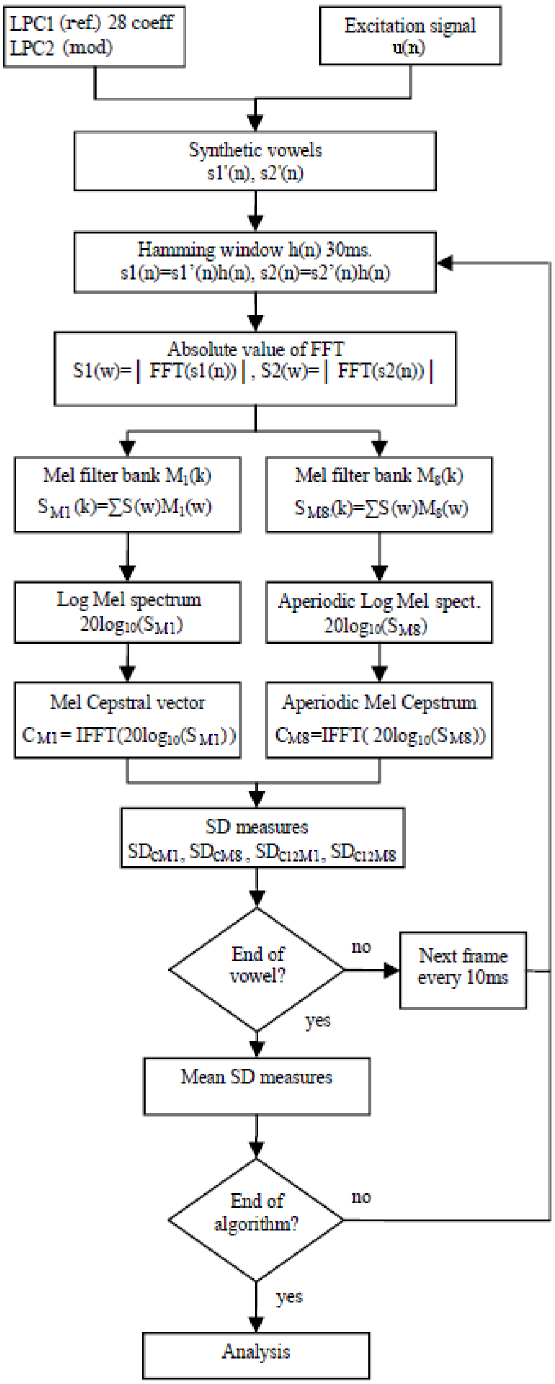

Fig. 2. Algorithm

For cepstral based objective measures, the analysis of synthetic vowels is performed on frames. Signals are sampled at $f_s = 16\,kHz$ and then analysed by using a 30 ms Hamming time window and 10 ms frame period.

$$s_p(n) = w_{\text{hamming}}(n) \cdot s_{\text{orig}}(p \cdot P + n) \quad 0 \le n \le N-1 \tag{7}$$

Extracted signal segment $s_p(n)$ for p^{th} analysis frame is found by multiplying the sampled vowel signal $s_{\text{orig}}(n)$ (reference or modified), with a Hamming window w_{hamming} of length N, where P is the frame period expressed in samples (10 ms $\cdot f_s$). The first stage of the MFCC analysis is computation of the spectrum $S(l)$ for each frame $s_p(n)$ of the signal by using N-point FFT.

$$S(l) = \sum_{n=0}^{N-1} s_p(n) \cdot e^{-2\pi inl/N} \quad 0 \le l \le N-1 \tag{8}$$

The Magnitude Spectrum is binned by $K + 1$ triangular filters whose central frequencies are equally spaced on the mel frequency scale. Since $S(l)$ is the signal spectrum above the linear frequency domain, the mel filter bank spectrum must also be converted to the linear [Hz] domain to perform the actual binning.

$$S_M(k) = 10 \cdot log_{10} \sum_{i=1}^{N/2} (|S(l)| \cdot M(k,l))^2 \quad 0 \le k \le K \tag{9}$$

$$S_M(-k) = S_M(k) \quad 1 \le k \le K \tag{10}$$

$M(k,l)$ is spectrum of the k-th mel filter mapped to the linear frequency domain and evaluated at the frequency of the l_{th} DFT bin. Taking the IFFT of the log mel spectrum $S_M(k)$, $-K \le k \le K$ gives the final MFCC vector.

$$c_M(n) = \frac{1}{2K+1} \sum_{k=-K}^{K} S_M(k) \cdot e^{2\pi ink/(2K+1)} \quad 0 \le n \le K \tag{11}$$

Log mel spectrum is a real and symmetric sequence that allows us to use the DCT instead of IFFT for calculation of the MFCC. The resulting MFCC vector is real and symmetric, thus having only K+1 unique samples. Frequency response of the analysis filter bank is chosen to mimic the process of human perception of speech, related to the frequency selective processing of audio signals performed along the basilar membrane in the cochlea of the inner ear. This filter bank is parameterized with the number of channels (K+1) and their respective bandwidth, expressed in mels. Increasing the number of channels improves the frequency selectivity of the analysis. Ideally, the central frequency of a triangular analysis filter can be treated as a continuous variable and the resulting cepstrum can be computed using the inverse discrete time Fourier transform (IDTFT instead of IDFT), giving an infinite symmetric aperiodic cepstrum sequence:

$$C_{\text{RealMel}}(n) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_{\text{LogMel}}(\omega) \cdot e^{i\omega n} d\omega \tag{12}$$

Filter bank with a finite number of channels actually samples the continuous mel-spectrum of the signal, what introduces aliasing in the cepstral domain. Therefore, the sequence $c_M(n)$ represents a periodic extension of the aperiodic sequence $C_{\text{RealMel}}(m)$ with the period of 2K+1. Described aliasing introduces ambiguity into signal representation since several different continuous mel-spectrums are mapped to the same periodic sequence $c_M(n)$. Aliasing primarily affects the cepstral samples on indices k close to K, but for certain “peaky” mel-spectra aliasing can also change the low-time cepstral samples that are used in objective distance measures.

In our research we have compared four different types of mel-cepstral based objective measures depending on the feature vector used for calculation. We have used either the aperiodic or periodic mel cepstral vectors, both in two different lengths, the full-length and truncated to the first 12 coefficients (what is a typical choice for real-word applications). The periodic mel cepstral vector is calculated by using the mel filter bank with an overlap coefficient 1 (Figure 3, red channels), such that edges of triangular channels exactly coincide with the centres of neighbouring channels. Since the aperiodic cepstral sequence indeed has an infinite length, it was approximated with a mel filter bank having an overlap coefficient of 8 (Figure 3, red and blue channels). In this way, the continuous mel-spectrum was sampled on an 8-times denser grid, but using the filters with identical mel-bandwidth as for the unit-overlap case. The resulting "approximately aperiodic" cepstral vector has a full length of 8K+1 samples (k = 0, ... 8K), thus reducing the effect of time-aliasing.

Two configurations of the mel filter bank are denoted with M_1 or M_8. As a measure of distance between the two vowels for one particular frame, we used RMS difference of the Log spectra above the mel-frequency scale.

$$\text{SD}_{\text{M1}} = \sqrt{\frac{1}{2K+1} \sum_{k=-K}^{K} (S_{\text{M1ref}}(k) - S_{\text{M1mod}}(k))^2} \tag{13}$$

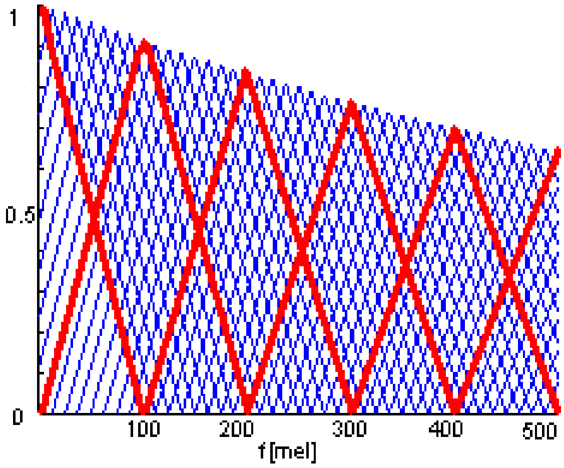

Fig. 3. Mel filter bank configurations

$$\mathrm{SD}_{\mathrm{M8}}=\sqrt{\frac{1}{2K+1}\sum_{k=-K}^{K}\left(S_{\mathrm{M8ref}}(k)-S_{\mathrm{M8mod}}(k)\right)^2} \tag{14}$$

Where S_{M1ref} and S_{M8ref} are log spectrums of the reference signal and S_{M1mod} and S_{M8mod} are log spectrums of the modified signal. According to Parseval equation:

$$\sum_{k=-K}^{K}|c_M(k)|^2=\frac{1}{2K+1}\sum_{k=-K}^{K}|S_M(k)|^2 \tag{15}$$

the MFCC based SD measure can be efficiently calculated from the cepstral coefficients. Coefficient $c_M(0)$ represents the average signal log-level that is related to the signal energy and because of that it is usually discarded from the distance measure. Finally, $\mathrm{SD}_{\mathrm{cM1}}$ and $\mathrm{SD}_{\mathrm{cM8}}$ distances can be rewritten using mel cepstral coefficients.

$$\begin{aligned}\mathrm{SD}_{\mathrm{cM1}}&=\sqrt{\sum_{k=-K,k\neq 0}^{K}\left(c_{\mathrm{M1ref}}(k)-c_{\mathrm{M1mod}}(k)\right)^2}=\\&=\sqrt{2\cdot\sum_{k=1}^{K}\left(c_{\mathrm{M1ref}}(k)-c_{\mathrm{M1mod}}(k)\right)^2}\end{aligned} \tag{16}$$

$$\mathrm{SD}_{\mathrm{cM8}}=\sqrt{2\cdot\sum_{k=1}^{K}\left(c_{\mathrm{M8ref}}(k)-c_{\mathrm{M8mod}}(k)\right)^2} \tag{17}$$

For many practical speech and speaker recognition tasks, cepstral coefficients sequences $c_M(k)$ are truncated to a length K_{fix} that is shorter then the actual number of filter bank channels K. It is assumed that most of the perceptually relevant information is retained in the first K_{fix} coefficients. A very common choice for this length is $K_{\mathrm{fix}}=12$. The corresponding distances $\mathrm{SD}_{\mathrm{c12}}$ between two such cepstral vectors can be found by truncating the sum in equation (16) and (17) to the first K_{fix} coefficients.

$$\mathrm{SD}_{\mathrm{c12M1}}=\sqrt{2\cdot\sum_{k=1}^{12}\left(c_{\mathrm{M1ref}}(k)-c_{\mathrm{M1mod}}(k)\right)^2} \tag{18}$$

$$\mathrm{SD}_{\mathrm{c12M8}}=\sqrt{2\cdot\sum_{k=1}^{12}\left(c_{\mathrm{M8ref}}(k)-c_{\mathrm{M8mod}}(k)\right)^2} \tag{19}$$

Note that distance $\mathrm{SD}_{\mathrm{c12}}$ between two truncated cepstral vectors is always smaller or equal to the distance SD_c of corresponding cepstrum sequences of the full length K.

4 EXPERIMENTS, RESULTS, DISCUSSION

4.1 Quality of the objective measure

In our first experiment we compared five different objective measures with subjective perceptual judgement. As we mentioned before, objective measures that were used are: SD distances computed from LPC parameters (equation (5)) and mel-cepstral based distances computed from four different mel cepstral vectors (equations (16) - (19)). Figures 4 to 7 show relation between LPC based measure (on the y-axis) and each one of the four mel cepstral based measures. Type of the symbols corresponds to the results of the subjective test (table 1). Gray-scale shaded regions represent perceptually different areas according to the KNN-3 classification.

As we can see, none of the objective measures predicts the subjective perceptual space ideally. We can not draw a clear border between perceptually audible and inaudible areas for any of the objective measures. But quite notably, the MFCC SD measure based on the full length cepstral vector with mel filter bandwidth of 220 mels would be the best choice that is shown in the central plot in the figures 4 and 5, with nice grouping of perceptual results along the x-axis. However, there are three small outlier groups of samples which are 'spoiling' the ideal separation of perceptually segregated areas. Next, we will analyze these groups of samples in detail (marked with 1, 2 and 3 in Figure 8).

The first outlier group represents perceptually clearly distinguishable vowel pairs with very small value of the

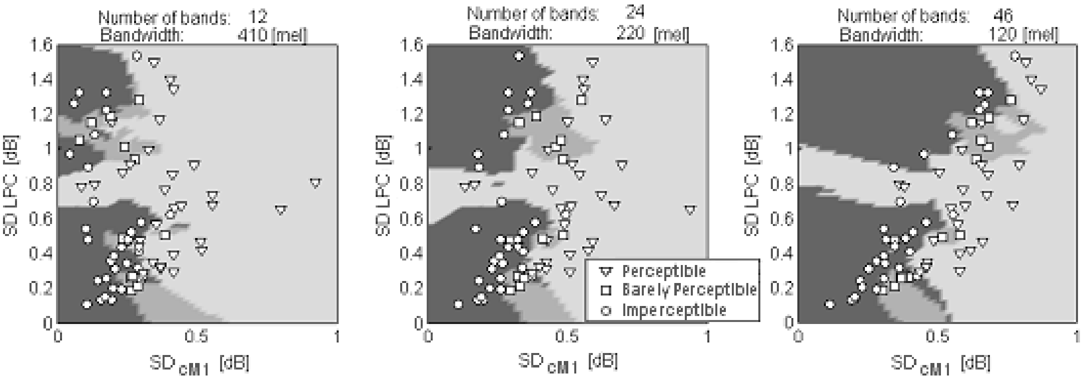

Fig. 4. Perceptual test graph, relation between LPC based measure SD_{LPC} and full lenght periodic mel cepstral based measure SD_{cM1}

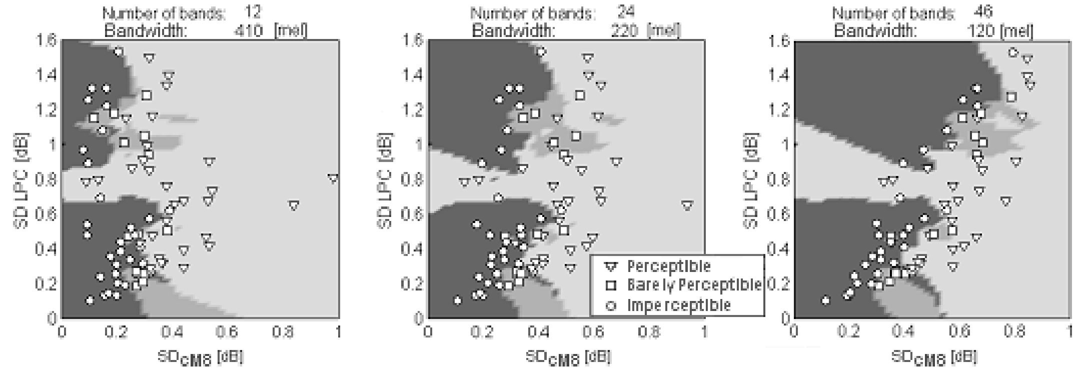

Fig. 5. Perceptual test graph, relation between LPC based measure SD_{LPC} and full lenght aperiodic mel cepstral based measure SD_{cM8}

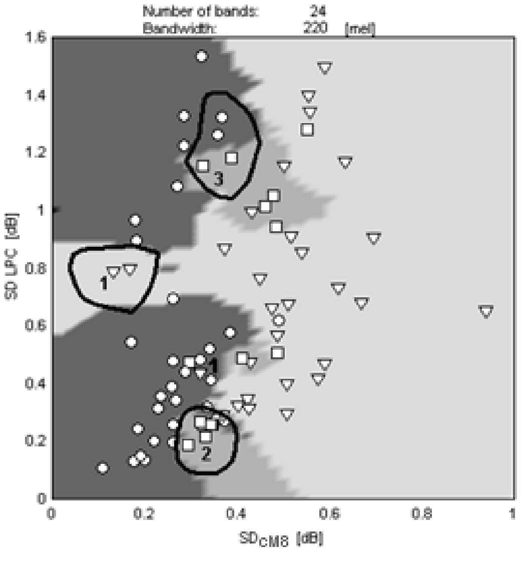

Fig. 8. Groups of samples 'spoiling' ideal classification

mel-cepstral distance measure. By analysing these samples in detail we noticed that the modified vowel in each of the pairs had a very sharp and narrow third or fourth formant (with bandwidth of less then 60 Hz). Such vowels sound unnatural, metallic, and therefore the difference can be easily noticed by the listener. The difference is also captured in the LPC based distance shown on the y-axis (of cca. 0.8 dB). However, since the mel filter channels are quite sparse and wide at frequencies corresponding to the third and fourth formant (F3 = 2.5 kHz and F4 = 3.5 kHz) the sharp change of the spectral envelope is overseen by mel-cepstral measure [6]. The actual formant bandwidths of speech sounds are typically of the order of 30 Hz to 70 Hz for formants below 2 kHz but increase for formants above 2 kHz due to various losses in the vocal tract. Higher formants have a typical bandwidth of around 250 Hz [19]. Because of that we can consider such modified vowels 'unnatural' and discard them from analysis, since after all the MFCC based measures are supposed to discriminate only between the natural speech sounds.

The second outlier group (Figure 8) represents pairs of vowels that are very close except in their first formant.

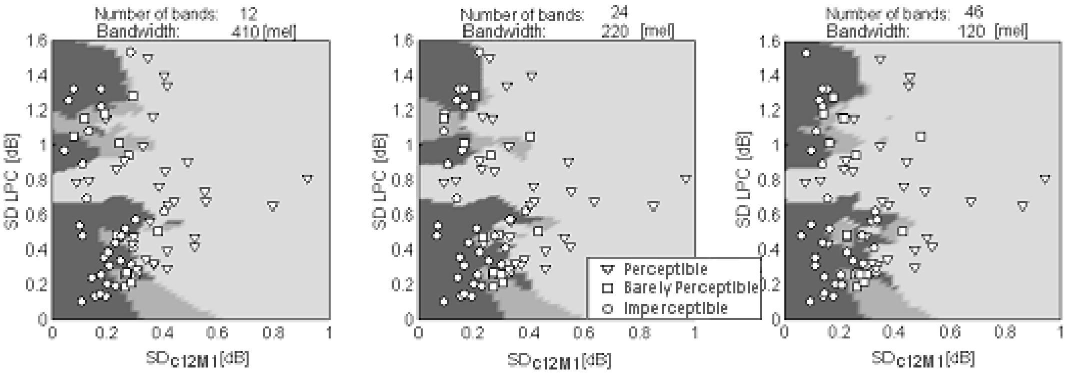

Fig. 6. Perceptual test graph, relation between LPC based measure SD_{LPC} and truncated periodic mel cepstral measure SD_{c12M1}

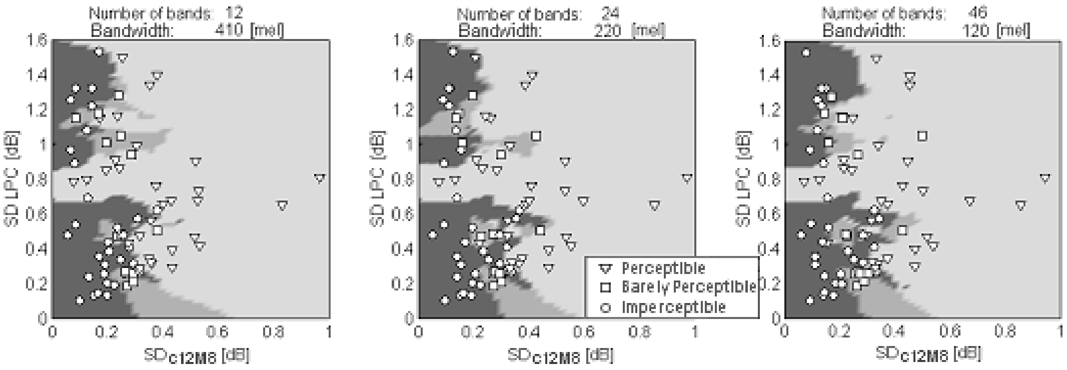

Fig. 7. Perceptual test graph, relation between LPC based measure SD_{LPC} and truncated aperiodic mel cepstral based measure SD_{c12M8}

Even though both objective measures between vowels give a very small distance (pairs are within inaudible area) perceptual difference is not completely inaudible. We can conclude that human auditory system is more sensitive to the first formant change than to any other formant, measured by the same objective measure [20]. Pairs of vowels from the third outlier group (Figure 8) have very large LPC based SD difference but rather small value of the mel-cepstral based measure. The reason for such discrepancy is that spectrums of these vowels are different on frequencies higher then 3.5 kHz especially in the valleys between higher formants. Since mel filter bands are sparse on these frequencies, spectral differences are poorly measured by mel cepstral measures. Such changes would be captured by MFCC measures better if we decrease the mel filter bandwidth (Figures 4 to 7). Even though SD LPC measure is very high for these vowel pairs, vowels are either indistinguishable or almost indistinguishable through listening. Consequently, we can conclude that human auditory system is much more sensitive to the changes at the lower frequencies, which clearly justifies the usage of the mel frequency scale.

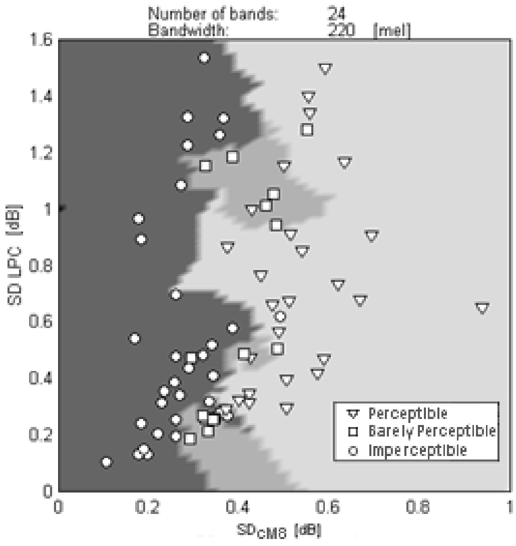

Fig. 9. Graph without 'unnatural' vowels

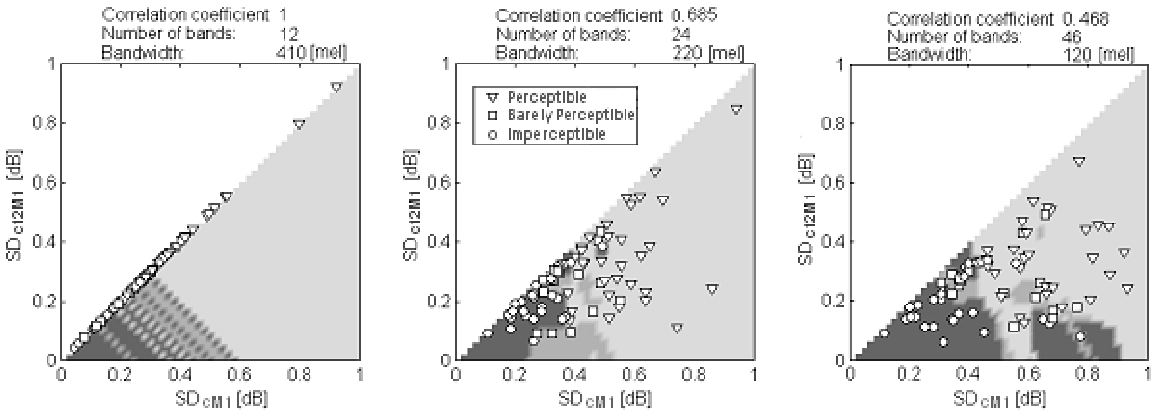

Fig. 10. Truncated versus full-length MFCC

Figure 9 compares the quality of the SD measures, cepstral versus LPC, without the 'unnatural' vowels of the first outlier group. There are three areas, two perceptually different areas "imperceptible" and "perceptible" and one undetermined area where samples of different classes are mixed. Our next goal was to determine the threshold of the cepstral based measure that separates these areas. The criteria we used to draw vertical borders between these areas is 95% purity of the area, meaning that 95% of the samples inside that area must belong to the same class. Such threshold analysis was performed for MFCC filter banks with fixed ovelap of 1 and bandwidth varying from 120 to 410 mels. Column 2 in table 2 gives the border (threshold) of the imperceptible area while column 3 represent the border of the perceptible area. Quality of the distance measure is estimated according to the width of the undetermined area or another word by the distance between perceptually different areas. Measure is considered better if undetermined area is as narrow as possible.

Table 2. Separation of the perceptualy different areas

Bandwidth [mel]	**Imperceptible [dB]**	**Perceptible [dB]**	**Undetermined area [dB]**
120	0.45	0.66	0.21
140	0.44	0.56	0.12
160	0.41	0.5	0.09
180	0.39	0.43	0.04
200	0.38	0.42	0.04
220	0.375	0.41	0.035
240	0.36	0.4	0.4
270	0.315	0.38	0.065
300	0.3	0.37	0.7
330	0.265	0.41	0.145
370	0.24	0.4	0.16
400	0.22	0.39	0.17
410	0.22	0.39	0.17

As we can see from table 2, filter banks with bandwidth from 180 to 240 mels give good objective cepstral measure with narrow undetermined regions. The best among them is the objective measure based on the full length cepstral vector calculated for the mel filter bandwidth of 220 mels. It gives accurate prediction of the perceptual distance between stationary vowel sounds (see figure 9 and table 2). Average spectral distortion of 0.4 - 0.5 dB represents a border between perceptually distinguishable areas. However, no clear border exist along the y-axis, showing that LPC based SD distance measured along the linear frequency scale is indeed a poor match to the human perception.

4.2 Truncated versus full-length MFCC

In digital speech analysis algorithms the most commonly used feature vector is the truncated mel cepstral vector MFCC. In our analysis, we used the typical truncation to the first 12 components without the coefficient c_0 which corresponds to the signal energy.

The second experiment investigates how such truncation affects the compliance of the objective measure with the perceptual difference between vowels. The experiment shows (figure 10) that correlation between the full length and truncated vectors depends on the number of filter channels. If $K = 12$, correlation is ideal since no information is discarded by truncation. However, for $K = 24$ or $K = 46$

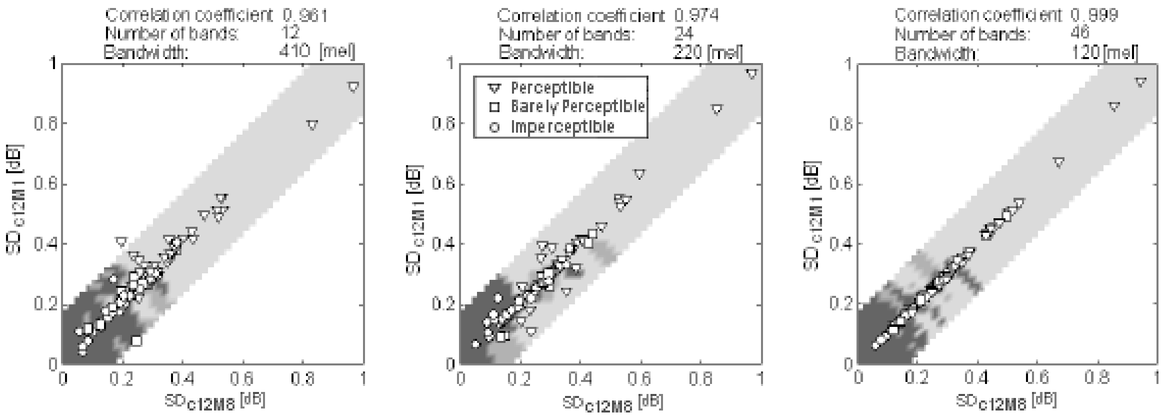

Fig. 12. Aperiodic versus periodic MFCC

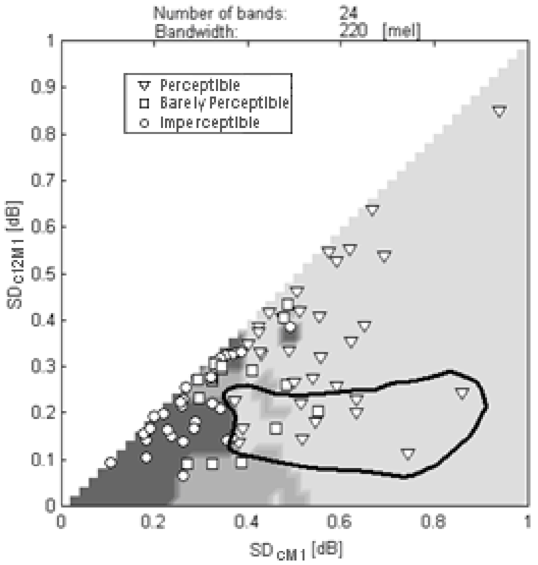

Fig. 11. Samples with small value of truncated MFCC based SD measure

correlation becomes much weaker. This experiment also shows that objective measure based on the full length vector predicts the subjective distance measure more effectively. Obviously, grouping of perceptually equal classes is much better along the x axis, then it is along the y-axis that corresponds to the truncated cepstral vector. Correlation between these two measures is also calculated using the Pearson product-moment correlation coefficient [21] (equation (20)):

$$r = \frac{N\sum_{i=1}^{N} X_i Y_i - \left(\sum_{i=1}^{N} X_i\right)\cdot\left(\sum_{i=1}^{N} Y_i\right)}{\sqrt{(N\sum_{i=1}^{N} X_i^2 - (\sum_{i=1}^{N} X_i)^2)\cdot(N\sum_{i=1}^{N} Y_i^2 - (\sum_{i=1}^{N} Y_i)^2)}} \tag{20}$$

Where N is number of samples, X_i is x coordinate of the i_{th} sample and Y_i is y coordinate of the i_{th} sample.

Again, we will analyse some extreme samples, circled in figure 11 that have large disagreement between the objective measures. For these vowel pairs, the mel-cepstral measure based on the truncated vector fails to detect pairs that are clearly distinguishable through listening. We found out that in these cases vowels spectrums are identical in their first two formants but different in the 3rd and especially in the 4th formant. It suggests that distance measure based on the truncated mel cepstral vector can be successfully used in speech recognition applications. For ASR the information about the first two formants is the most important since it conveys most of the speech information. However, usage of such simplified measure in speaker recognition application would not be fully justified, since it is exactly the higher formant regions that are relevant for differentiating between speakers.

4.3 Aliasing

The third experiment analyses the influence of aliasing in cepstral domain on the mel-cepstral based objective measure caused by sampling of the continuous mel spectrum due to a finite number of analysis channels.

For that purpose we compared objective measures based on aperiodic and periodic mel cepstral vectors. Ideally the aperiodic mel cepstral vector is obtained by using mel filter bank with infinite number of channels and thus the mel-spectrum becomes a continuous function of the filter central band frequency (12). Periodic mel cepstral vector (regular MFCC) is obtained by using a finite number of channels, or equivalently by periodic extension of the aperiodic sequence with the period of 2K+1 (11). In our experiment the aperiodic mel cepstral vector was not actually infinite, but approximated using a mel filter bank with 8-times overlap of neighbouring channels (figure 3). Thus the resulting vector is 8 times longer then the periodic mel cepstral vector.

From figure 12 it is visible that correlation coefficient is quite close to 1 regardless of the filter parameters values. That suggests that for truncated mel-cesptral distances the effect of aliasing is relatively insignificant. The results also demonstrate that aliasing is reduced with increasing the number of analysis channels.

For $K = 46$ the truncated cepstral distance derived from the periodic cepstrum is almost identical to the one derived from the aperiodic one. Such behaviour is in accordance with intuition, since with rapidly decaying cepstrum and with the long period of periodic extension (2K+1), the cepstrum samples around the origin that are used in the truncated vector are hardly affected by aliasing. Interestingly, although correlation coefficient is very close to 1, even for $K = 12$ or $K = 24$, there are always certain samples with significantly different values of objective measures based on periodic or aperiodic mel-cepstral vector. Such samples are the most interesting for detailed aliasing analysis. It was found that the sample with significant aliasing influence on objective measures for one set of filter parameters does not exhibit the same influence for even slightly modified set of parameters. We have analysed the influence of the filter bandwidth and filter overlap coefficient on the aliasing.

In our first experiment the number of filter channels was fixed to 24. Two filters overlap coefficients were used: 1 (periodic cepstrum) and 8 (aperiodic cepstrum). Bandwidth of 220 mels ensures exact coverage of the whole spectrum using 24 filters. In order to check the influence of channel bandwidth on aliasing, the bandwidth was varied slightly, in the range from 210 to 230 mels. Of course, with the narrower bandwidth (210 mel), the last channel is positioned cca. 5% below the nominal centre mel frequency, while for the wider case (230 mel) it goes 5% beyond the Nyquist frequency. Described procedure contracts or expands the frequency response of the analysis filter bank, thus slightly realigning the filters' centre frequencies relative to the formant structure of the vowel. As it is shown in figure 13, even such small change affects both objective measures SD_{c12M1} and SD_{c12M8}.The distance from diagonal correlation line determines the aliasing impact. The difference between the two measures can be attributed to the aliasing, while variations of individual measure to the bandwidth change can be attributed to the modified sampling grid of the signal spectrum.

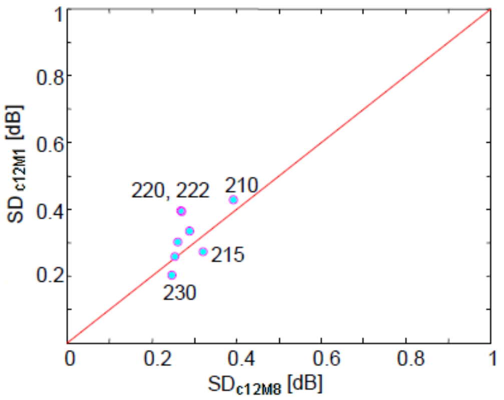

Fig. 13. SD measures for different mel filter bandwidths

In our second experiment we have analysed the relation between objective measures calculated for different overlap factors of the neighbouring analysis channels. Mel filter bandwidth was fixed to 220 mels while number of filter channels was varied, according to chosen overlap factor, to keep the spectrum fully covered with the span of the analysis bank. We have chosen 17 vowel pairs with significant aliasing impact. Correlation between aperiodic cepstral vector (overlap coefficient 8) and periodic vector (overlap coefficient from 2/3 to 4) is calculated using (20). As we can see (table III), the influence of aliasing decreases rapidly as filter overlap coefficient increases. With overlap factor of two, the difference between aperiodic and periodic distance measures become almost negligible, even for critical vowel pairs.

Table 3. Overlap - Correlation

Overlap cofficient	8	4	2	1	2/3
Corelation coefficient	1	0,9997	0,9991	0,7238	0,6098

5 CONCLUSIONS

Perceptual significance of the mel cepstral measures was evaluated by comparing spectral distance of the LPC models with several objective distance measures derived from the mel cepstral coefficients. The measure used was SD RMS distortion measure. Feature vectors were derived from either periodic or aperiodic mel-cepstrum either as full length vectors or truncated to the first 12 coefficients. Vectors were calculated for different values of the mel filter bank parameters, i.e. number of channels and bandwidth / overlap factor. The first experiment examined compatibility of the MFCC based distance measures with human perception for different values of analysis parameters. By varying mel-filter bank parameters we found that bank with 24 channels, with bandwidth of 220 mels (i.e. with overlap factor equal or higher than one) gives MFCC feature vectors which are very good perceptually compliant representation of stationary vowels. For this kind of analysis, perceptual difference between vowels can be recognised if the full-length mel cepstral SD RMS distance is higher than 0.4 - 0.5 dB. There are certain exceptions to this compliance which were analysed in detail. Modified vowels with sharp and narrow 3rd or 4th formant are easily perceptually distinguishable from the original vowels, but mel cepstral distance does not capture this difference. Such vowels can be deemed unnatural and were discarded from the analysis. However, small difference between vowels' first formants and/or significant spectral differences of the higher frequencies regions were also clearly audible in subjective listening test, but were poorly measured by the MFCC based measures. In the first case, the reason may be very high sensitive of the human auditory system for low frequency modifications, around the frequency of the first formant. In the second case, mel-cepstral objective measure fails to measure rather significant spectral differences in higher frequencies regions due to a fact that mel filterbank channels are sparse and wide in that frequencies range when mapped to the linear frequency domain.

The second experiment examined the relation between full length and truncated MFCC measures and their compliance to the subjective measurements. The experiment showed that measure based on full-length vector always gives better compliance. By analysing cases in which truncated feature vector fails to measure difference between vowels we found that truncated vector models the first two formants adequately, but not the 3rd and the 4th formant. That suggests that truncated MFCC vector would be a sufficiently good choice for speech recognition applications but its usage may be arguable for speaker recognition applications.

The third experiment analysed the impact of aliasing in cepstral domain on the truncated cepstral based distance measures. The results showed high correlation of SD distances calculated from aperiodic and periodic mel cepstrum, leading to conclusion that the impact of aliasing is generally minor especially if the number of analysis channels is high (e.g. K=46). There are rare exceptions when aliasing was indeed observed and which were used to examine the relation between mel filter bank parameters and aliasing in detail. It is found that changes in the filter overlap factor (as long it is one or higher) do not affect level of aliasing significantly. However, variation of the filter bandwidth directly affects the level of aliasing. Aliasing definitely introduces ambiguity into differentiation of speech sounds, but there is still a disputable question whether similar ambiguity also exists for the human perception as well. If this is really the case, then aliasing is perfectly acceptable, as long as it closely mimics the process related to human perception of speech sounds. Detailed investigation of this problem is the topic of the ongoing research.

REFERENCES

[1] A. Rix, J. Beerends, D. Kim, P. Kroon, and O.Ghitza, "Objective assessment of speech and audio quality, technology and applications," in *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 1890–1901, 2006.

[2] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," in *IEEE Journal on selected areas in Communications*, vol. 10, pp. 819–829, 1992.

[3] A. Rix, M. Hollier, A. Hekstra, and J. Beerends, "Perceptual evaluation of speech quality (pesq), the new itu standard for end-to-end speech quality assessment. part i - time alignment," in *J. Audio Eng. Soc.*, pp. 755–764, 2002.

[4] A. Rix, M. Hollier, A. Hekstra, and J. Beerends, "Perceptual evaluation of speech quality (pesq), the new itu standard for end-to-end speech quality assessment. part ii - psychoacoustic model," in *J. Audio Eng. Soc.*, pp. 765–778, 2002.

[5] A. Takahashi, A. Kurashima, and H. Yoshino, "Objective assessment methodology for estimating conversational quality in voip," in *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 1984–1993, 2006.

[6] S. Umesh and R. Sinha, "A study of filter bank smoothing in mfcc features for recognition of children's speech," in *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, pp. 2418–2430, 2007.

[7] S. Voran, "Objective estimation of perceived speech quality—part ii: Evaluation of the measuring normalizing block technique," in *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 383–390, 1999.

[8] D. O'Shaughnessy, "Interacting with computers by voice: automatic speech recognition and synthesis," in *Proc. IEEE*, vol. 91, pp. 1272–1305, 2003.

[9] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 357–366, 1980.

[10] L. Rabiner and B. Juang, *Fundamentals of speech recognition.* Prentice-Hall Inc., 1993.

[11] P. Aleksic and A. Katsaggelos, "Audio-visual biometric," in *Proceeding of the IEEE*, vol. 94, pp. 2025–2044, 2006.

[12] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, “Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006,” in *IEEE Trans., Audio, Speech and Language Processing*, 2007.

[13] D. Chazan, R. Hoory, G. Cohen, and M. Zibulski, “Speech reconstruction from mel frequency cepstral coefficients and pitch frequencyl,” in *Proceeding ICASSP*, 2000.

[14] B. Milner and X. Shao, “Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model,” in *In ICSLP-2002*, pp. 2421–2424, 2002.

[15] L. Rabiner and R. Schafer, *Digital processing of speech signals*. Englewood Cliffs, New Jersey: Prentice-Hall Inc., 1978.

[16] F. Itakura, “Line spectrum representation of linear predictive coefficients of speech signals,” in *J. Acoust. Soc. Amer.*, vol. 57, p. S35, 1975.

[17] J. C. Jr, “Speaker recognition: a tutorial,” in *Proceeding of the IEEE*, vol. 85, pp. 1437–1462, 1997.

[18] T. Cover and P. Hart, “Nearest neighbor pattern classification,” in *IEEE Trans. Information*, 1967.

[19] G. Fant, “Vocal tract wall effects, losses, and resonance bandwidths,” in *STL-QPSR, no. 2–3*, pp. 28–58, 1975.

[20] G. Peterson and H. Barney, “Control methods used in a study of the vowels,” in *J.Acoust.Soc.Am.*, vol. 24, pp. 175–184, 1952.

[21] W. Yang, *Enhanced modified bark spectral distorsion (EMBSD): an objective speech quality measure based on audible distortion and cognition model*. PhD thesis, Temple University, 1999.

Antonio Vasilijević received his Dipl. ing. degree in Electrical Engineering from University of Sarajevo in 1993 and M.S. degree from University of Zagreb in 2008. He is currently researcher at the Laboratory for Underwater Systems and Technologies (LaBUST), Faculty of Electrical Engineering and Computing, University of Zagreb. His research interests covers: marine and underwater technologies, augmented reality, acoustics, signal analysis and processing, pattern identification, recognition and classification.

Davor Petrinović (M’96) was born in 1965 in Croatia. He received the Dipl. ing. degree in Electrical Engineering from University Zagreb, Faculty of Electrical Engineering in 1988 (today, Faculty of Electrical Engineering and Computing). He received M.Sc. and Dr.Sc. degree in the field of electrical engineering, in 1996. and 1999. respectively, from the same institution. He was appointed Full Professor in 2010, at the Department of Electronic Systems and Information Processing, Faculty of EE&C, Uni. Zagreb. He was a Fulbright post. doc. scholar in 2000/01 at SCL Laboratory, UC Santa Barbara, USA and a visiting researcher at Sound and Image Processing Lab, School of EE, KTH, Stockholm Sweden in 2005/06. His current research interests include signal processing and speech and audio modeling, processing and coding. He is currently the Vice Dean for Accademic Affairs.

AUTHORS’ ADDRESSES

Antonio Vasilijević
Davor Petrinović
Faculty of Electrical Engineering and Computing,
University of Zagreb,
Unska 3, 10000 Zagreb, Croatia
email: antonio.vasilijevic@fer.hr, davor.petrinovic@fer.hr

Received: 2010-01-25
Accepted: 2011-02-01