DISCONTINUITY DETECTION IN CONCATENATED SPEECH SYNTHESIS BASED ON NONLINEAR SPEECH ANALYSIS

Yannis Pantazis, Yannis Stylianou

University of Crete, Computer Science Department,
Heraklion Crete, Greece
email: {pantazis, yannis}@csd.uoc.gr

Esther Klabbers

Center for Spoken Language Understanding,
OGI School of Science & Engineering,
OHSU, Beaverton, Oregon
email: klabbers@cslu.ogi.edu

10.21437/Interspeech.2005-621

ABSTRACT

An objective distance measure which is able to predict audible discontinuity in concatenated speech synthesis systems is very important. Previous works were primarily based on features estimated by linear and/or stationary models of speech. In this paper, we introduce two nonlinear approaches for the detection of discontinuity. The first method is based on a nonlinear harmonic model of speech while the second method is based on the demodulation of speech in an amplitude and a frequency component using the Teager energy operator. Fisher's linear discriminant was used for the separation of signals with audible discontinuity from those perceived as continuous. When we combined the two methods using Fisher's linear discriminant a detection rate of 56.5% was achieved which is an 90% improvement over previously published results on the same database.

1. INTRODUCTION

In many modern text-to-speech (TTS) synthesis systems, synthetic speech is produced by concatenating speech segments selected from a large inventory [1], [2], [3], [4]. In these inventories, there are many instances for each speech segment (referred to as unit) with various prosodic and spectral characteristics. For high-quality and natural-sounding speech synthesis, units have to be selected in an optimum way. The selection process uses a combination of two costs attributed to each candidate unit. The first cost, which is called the target cost, expresses the closeness between the context of the target and that of the candidate unit. It is calculated as a weighted sum of differences between prosodic and phonetic parameters. The other cost which is called concatenation cost refers to how well adjacent units can be joined. It is calculated as a weighted sum of differences between fundamental frequency, spectral mismatches, energy, etc. Optimum unit selection is achieved by a Viterbi search for the lowest total cost path through the lattice of candidate units. Between these two costs, the concatenation cost is usually considered to be the most important one in the selection process.

Many current studies have as focus to define a concatenation distance measure that would be able to predict audible discontinuity. Such an objective measure should be highly correlated with human perception results from subjective tasks where discontinuity in the concatenation of units was considered by humans to be audible. Wouters and Macon [5] found that the Euclidean distance on mel-scale LPC-based cepstral coefficients performed well. Klabbers and Veldhuis [6] found that the best predictor of audible discontinuity was the Kullback-Leibler distance on LPC power spectra. Stylianou and Syrdal [7] showed that Kullback-Leibler distance on FFT-based power spectra was the best predictor. Donovan [8] proposed Mahalanobis distance between perceptual cepstral parameters employing decision trees. Vepa et al. [9] used Kalman filtering for the evaluation of join costs. Bellegarda [10] proposed an SVD-based Fourier analysis for assigning concatenation costs. Most of these studies were phoneme specific and only a few of them were phoneme independent. Phoneme specific approaches provide better results compared to phoneme independent approaches. This is expected since in the former case the search space is smaller compared to the space generated in the phoneme independent analysis case. However, even for these phoneme specific approaches the prediction score cannot be considered to be sufficiently high. Moreover, these studies were conducted on different databases. Thus, it is not possible to make direct comparisons between features and methods that were used in different studies and draw useful conclusions about them. Last but not least, most of previous approaches consider the signals to be stationary. Therefore, estimation of features was mainly based on stationary signal representations (parametric or not). However, when two non-contiguous speech segments are concatenated together the final signal is expected to present non-stationary characteristics (even if *linear* phase mismatches have been reduced). Linear models cannot, then, accurately represent the generated non-stationary signal. We believe that these fast changes in the statistical properties of the signal are detected as discontinuity.

In this paper, two new sets of features for detecting audible discontinuity and a new discrimination function are introduced. In order to increase the detection rate we suggest using features from a nonlinear speech model and from a nonlinear speech analysis algorithm. The first set of features are obtained by modeling the speech signal as a sum of harmonics with time varying complex amplitude [11]. The second set of features is based on a technique trying to decompose speech signals into AM and FM components [12]. We propose to work in the initial dimension of the estimated data by applying discriminant functions rather than to work with a reduced dimension (e.g., using a simple Euclidean distance). We suggest using Fisher's linear discriminant [13] as a discrimination function. The evaluation of the objective distance measures was conducted on the database created by Klabbers and Veldhuis [14]. Therefore we are able to compare different approaches (the one proposed herein and the other described by Klabbers and Veldhuis)

Esther Klabbers' contribution is supported by NSF grant 0313383: "Objective Methods for Predicting and Optimizing Synthetic Speech Quality"

on the same database.

The paper is organized as follows. In section 2 the extraction of the two sets of parameters is presented while in section 3 Fisher's linear discriminant is quickly reviewed. Section 4, describes the speech database used and the listening experiment. Results from the evaluation of various distance measures are presented in section 5. A summary on the derived results as well as future work concludes the paper.

2. NEW FEATURE SET

In previous published work on this subject, speech was considered as a stationary process around the concatenation point. Hence, the techniques used for the extraction of the feature set did not take into account any dynamic information of the speech signal. But experimental work provided evidence that even in continuous speech, resonances can change rapidly within a few - even a single- speech periods [15]. Therefore, in an attempt to incorporate dynamic information in the decision whether or not there is an audible discontinuity, a set of nonlinear features are extracted from the concatenated speech signal.

2.1. A Nonlinear Harmonic Model

The first technique is based on a nonlinear harmonic representation of speech signals [11]. The model assumes the speech signal to be composed of a periodic signal, h[n], which is designated as sums of harmonically related sinusoids

$$h[n] = \sum_{k=-L(n_i)}^{L(n_i)} A_k[n] e^{j2\pi k f_0(n_i)(n-n_i)} \tag{1}$$

where $L(n_i)$ denotes the number of harmonics at $n = n_i$, $f_0(n_i)$ denotes the fundamental frequency at n = n_i, while

$$A_k[n] = a_k(n_i) + (n - n_i) b_k(n_i) \tag{2}$$

where $a_k(n_i)$ and $b_k(n_i)$ are assumed to be *complex* numbers which denote the amplitude of the k^{th} harmonic and the first derivative(slope), respectively.

The size of analysis window is two pitch periods. First, the current fundamental frequency, $f_0(n_i)$, is evaluated from the autocorrelation function of the speech signal around the concatenation point. For an efficient speech representation the whole spectrum must be considered; therefore, the number of harmonics, $L(n_i)$, must be computed as $L(n_i) = \lfloor \frac{f_s}{2 f_0(n_i)} \rfloor$ where f_s denotes the sampling frequency and $\lfloor \rfloor$ denotes the floor operator. However, since the primary goal here is not the signal representation but the detection of audible discontinuity at concatenation points, a subset of frequencies may be used. We expect that lower frequencies are more important for our task. Therefore, the number of features from the harmonic model may be reduced, also decreasing the complexity of the detection process. In this paper only the first $4000Hz$ were taken into account.

The unknown complex amplitudes (eq. (2)) are estimated by minimizing a weighted time-domain least-squares criterion with respect to $a_k(n_i)$ and $b_k(n_i)$,

$$\epsilon = \sum_{n=n_i-T_0}^{n=n_i+T_0} w^2[n](s[n] - h[n])^2 \tag{3}$$

where $s[n]$ denotes the original speech signal, $h[n]$ denotes the harmonic signal to estimate, $w[n]$ denotes the weighted window (which is a Hanning window) and T_0 denotes the local fundamental period $(f_s/f_0(n_i))$, in samples.

2.2. AM&FM Components

The Teager-Kaiser (TK) energy operator was defined in [16] to be

$$\Psi\{x[n]\} = x^2[n] - x[n-1]x[n+1] \tag{4}$$

Based on this operator Maragos et al. [12] have developed the Discrete Energy Separation Algorithm (DESA) for separating an AM-FM modulated signal into its components. One version of DESA is described by the following equations:

$$G[n] = 1 - \frac{\Psi\{y[n]\} + \Psi\{y[n+1]\}}{4\Psi\{x[n]\}} \tag{5}$$

$$\Omega[n] \approx \arccos(G[n]) \tag{6}$$

$$|a[n]| \approx \sqrt{\frac{\Psi\{x[n]\}}{1 - G^2[n]}} \tag{7}$$

where $y[n] = x[n] - x[n-1]$, $\Omega[n]$ is the instantaneous frequency and $a[n]$ is the instantaneous amplitude.

One application of DESA in speech analysis is the separation of a signal around a resonance in an amplitude and a frequency component [17]. The extraction of a single resonance is done by bandpass filtering the speech signal with a Gabor filter with impulse response defined by

$$h_G[n] = \exp(-b^2 n^2) \cos(\Omega_c n) \tag{8}$$

where b and Ω_c control the bandwidth and the central frequency of the filter respectively. The size of analysis window was set to 300 samples (approximately 20ms).

A filterbank of twenty Gabor filters was used. The impulse response of the filter was 150 samples long and the value of b was selected to be 250; hence the bandwidth of each filter was approximately 425Hz. In order to cover most of the spectrum and since the sampling frequency(F_s) of the recordings was 16kHz, the central frequencies of the filters were uniformly distributed between 250Hz and 5000Hz.

2.3. Features

The features were extracted from a database containing synthetic test words. Each word consists of two parts (therefore, there is only one concatenation point per word): a left part and a right part. From each part a set of features was estimated. Many options may be considered for the comparison of these features. We present those that gave high detection rates while at the same time, they have an intuitive meaning. For instance, since the features estimated by the harmonic model are complex numbers, the absolute of their complex difference is equivalent to the Euclidean distance between two points on the complex plane. For the second set of parameters, the AM features are defined by a metric measured as the l_1 norm (sum of the absolute differences) between the AM components estimated for the left and right part. The same metric was used for the FM features.

3. DISCRIMINATION FUNCTION

Until now, research on predicting audible discontinuity in concatenated speech synthesis was concentrated on detecting the right features and an appropriate distance measure for this task. In our approach, we construct a feature vector - hence a feature space - for each speech signal instead of finding a distance measure. Then, we define two classes: one for perceptually discontinuous signals and another for signals that were detected to be continuous. Then, statistical methods may be applied for an efficient separation of the two classes. We suggest the use of Fisher's linear discriminant. An advantage of using Fisher's linear discriminant for the separation of the two classes is its simplicity, as well as, its direct comparison with distances used so far.

3.1. Fisher's Linear Discriminant

Suppose that we have a set of N d-dimensional samples $\mathbf{x_1}$,...,$\mathbf{x_N}$, N_0 samples be in the subset D_0 and N_1 samples be in the subset D_1. If we form a linear combination of the elements of $\mathbf{x}$, we obtain the scalar dot product

$$y = \mathbf{w^T x} \tag{9}$$

and a corresponding set of N samples y_1,...,y_N that is divided into the subsets Y_0 and Y_1. This is equivalent to form a hyperplane in d-space which is orthogonal to $\mathbf{w}$.

The direction of $\mathbf{w}$, important for maximum separation, is given by

$$\mathbf{w} = \mathbf{S_W^{-1}}(\mathbf{m_0} - \mathbf{m_1}) \tag{10}$$

where

$$\mathbf{S_W} = \sum_{\mathbf{i=0}}^{\mathbf{1}} \sum_{\mathbf{x}\in D_i} (\mathbf{x} - \mathbf{m_i})(\mathbf{x} - \mathbf{m_i})^{\mathbf{T}} \tag{11}$$

and

$$\mathbf{m_i} = \frac{1}{\mathrm{N_i}} \sum_{\mathbf{x}\in D_i} \mathbf{x}\,, \qquad \mathrm{i} = 0, 1. \tag{12}$$

Since Fisher's linear discriminant projects feature vectors to a line it can also be viewed as an operator (FLD) which is defined by

$$FLD\{\mathbf{x}\} = \sum_{i=1}^{d} w_i x_i \tag{13}$$

where w_i are the elements of $\mathbf{w}$. If x_i are real positive numbers, this is a kind of weighted version of l_1 norm (weights can be negative numbers). According to this method, we are now able to combine features which are on a different scale.

3.2. Detection Scenario

In distance measures as well as in vector projection we deal with scalars. The evaluation of the distance measures was based on the detection rate, P_D, given a false alarm rate, P_{FA}. For each measure, y, two conditional probability density functions, $p(y|C_0)$ and $p(y|C_1)$ were computed depending on the results from the perceptual test; C_0, if the synthetic sentence was perceived as continuous, and C_1 if it was perceived as discontinuous by the listeners. Then the detection rate for that measure, y, is computed as:

$$P_D(\gamma) = \int_{\gamma}^{\infty} p(y|C_1)\, dy \tag{14}$$

where γ is estimated by:

$$P_{FA}(\gamma) = \int_{\gamma}^{\infty} p(y|C_0)\, dy = 0.05 \tag{15}$$

4. DATABASE AND LISTENING EXPERIMENT

In this section we briefly present the database as well as the listening experiment that was conducted.A more detailed description can be found in Klabbers et al. [6].

Five subjects with backgrounds in phycho-acoustics or phonetics participated in the listening experiment. The material was composed of 1449 C_iVC_j stimuli, which were constructed by concatenating diphones C_iV and VC_j excised from nonsence words of the form $C@CVC@$ (where C = consonant, V = vowel $\in$ /a/, /i/, and /u/, and @ = schwa). The recordings were made of a semi-professional female speaker and resampled to 16 kHz.

Preliminary tests showed that discontinuities and other effects in the surrounding consonants would overshadow the effects in the vowel. Hence the surrounding consonants were removed. In addition, the duration of the vowels was normalized to 200 ms and the signal power of the second diphone was scaled to equalize the level of both diphones in the boundary. The stimuli were randomized and the subjects were instructed to ignore the vowel quality and focus on the diphone transition. Their task was to make a binary decision about whether the transition was smooth (0) or discontinuous (1). The experiment was divided into six blocks, presented in three hourly sessions with a short break between two blocks. A transition was marked as discontinuous when the majority of the subjects (3 or more out of 5) perceived it as such.

5. RESULTS AND DISCUSSION

In many previous publications symmetric Kullback-Leibler (SKL) divergence has been shown to provide the highest correlation with human perception results [14] [7]. Using the same database as in [14] and without splitting the database in a training and a testing dataset, SKL on a smoothed magnitude spectrum computed by linear prediction coefficients has a detection rate of 30.90%. Please note that this score is for a phoneme-independent scenario. This result along with the results for the new features using FLD are presented in Table 1. The false alarm for all detection scores was set to 5%. FLD using amplitudes, a_k, of the harmonic model

Distance	Detection Rate (%)
SKL	30.90
a_k	39.57
b_k	32.61
a_k & b_k	46.52
AM	38.61
FM	19.66
AM & FM	49.40
a_k & b_k & AM & FM	56.35

Table 1: Detection Rates for a phoneme independent task

gave a detection rate of 39.57% while slopes, b_k, gave a rate of 32.61%. It turns out that amplitudes performed better than slopes.

However, both rates are higher than the rate obtained using SKL on a smoothed spectrum. The combination (by simple concatenation) of amplitudes and slopes increases the detection rate. A few remarks may be made here. First, the detection rate by the simple concatenation of features is not equal to the sum of the individual scores. This means that there is a correlation between these two parameters. On the other hand, by using the slopes as features a (relatively to other scores) high detection score is obtained. This shows the importance of slopes, or otherwise of the nonlinear features, for this task.
Regarding the features extracted by the AM & FM components, AM based features outperform FM based features. AM features perform approximately the same as the amplitudes a_k. By combining AM and FM features (simple concatenation) the score is higher than that obtained by the combination of amplitudes and slopes, showing that there is less correlation between these two features. Finally, by applying FLD on the whole set of features (Harmonic parameters, AM, and FM) a detection rate of 56.35% was obtained which is an 90% improvement over previously published results on the same database.
From the above results it is obvious that elimination of redundancy between features as well as a better fusion of them will result in higher detection rates.

The ROC (Receiver Operating Characteristic) curves for a_k, b_k, AM and FM are depicted in Figure 1. In the same figure, results using smooth spectrum and SKL divergence are also included for comparison purposes.

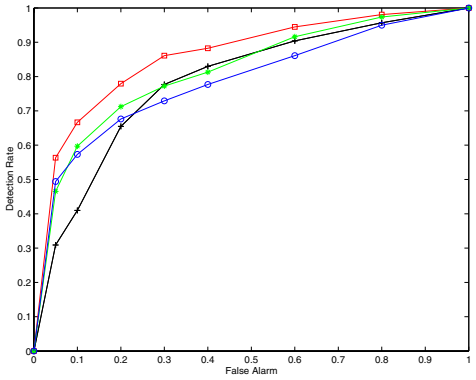

Figure 1: ROC: SKL (plus), a_k&b_k (star), AM&FM (circle) and a_k&b_k&AM&FM (square)

6. CONCLUSIONS

This paper introduced two new feature sets for the problem of detection audible discontinuity in concatenated speech synthesis. The first set of features were extracted from a nonlinear speech model which assumes speech signals as a sum of harmonic sinusoids. The second set of features was based on a method that decomposes speech signals into AM and FM components. Signals with audible discontinuity were separated from those without audible discontinuity by a hyperplane which was determined by Fisher's linear discriminant. A high detection rate (compared to previous published results on the same database) was obtained when the above features were combined. However, we expect that better results can be obtained by reducing redundancy between features, exploring better fusion strategies of the features and finally, using more sophisticated discrimination functions.

7. REFERENCES

[1] A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using large speech database. *Proc. IEEE Proc. ICASSP-1996*, pages 373–376, 1996.

[2] W. N. Campbell and A. Black. Prosody and the selection of source units for concatenative synthesis. In R. Van Santen, R.Sproat, J.Hirschberg, and J.Olive, editors, *Progress in Speech Synthesis*, pages 279–292. Springer Verlag, 1996.

[3] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. The AT&T Next-Gen TTS System. *137th meeting of the Acoustical Society of America*, 1999. http://www.research.att.com/projects/tts.

[4] G. Coorman, J. Fachrell, P. Rutten, and B.Van-Coile. Segment selection in the l&h realspeak laboratory tts system. *Proc. ICSLP 2000*, 2000.

[5] J. Wouters and M. Macon. Perceptual evaluation of distance measures for concatenative speech synthesis. *International Conference on Spoken Language Processing ICSLP 98*, pages 2747–2750, 1998.

[6] E. Klabbers and R. Veldhuis. Reducing audible spectral discontinuities. *IEEE Transactions on Speech and Audio Processing*, 9:39–51, Jan 2001.

[7] Y. Stylianou and A. Syrdal. Perceptual and objective detection of discontinuities in concatenative speech synthesis. *Proc. IEEE Proc. ICASSP-01*, 2001.

[8] Robert E. Donovan. A new distance measure for costing spectral discontinuities in concatenative speech synthesis. *The 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, 2001.

[9] J. Vepa and S. King. Kalman-filter based join cost for unit selection speech synthesis. *Eurospeech 2003*, 2003.

[10] Jerome R. Bellegarda. A novel discontinuity metric for unit selection text-to-speech synthesis. *5th ISCA Speech Synthesis Worksop*, pages 133–138, 2004.

[11] Yannis Stylianou. *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis, Ecole Nationale Supèrieure des Télécommunications, 1996.

[12] P. Maragos, J. Kaiser, and T. Quatieri. On separating amplitude from frequency modulations using energy operators. *Proc. IEEE Proc. ICASSP-92*, Mar 1992.

[13] R.O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, Inc, 2001.

[14] E. Klabbers and R. Veldhuis. On the reduction of concatenation artefacts in diphone synthesis. *International Conference on Spoken Language Processing ICSLP 98*, pages 1983–1986, 1998.

[15] H. M. Teager and S. M. Teager. Evidence for nonlinear sound production mechanism in the vocal tract. *Speech Production and Speech Modelling*, 55, Jul 1990.

[16] J. F. Kaiser. On a simple algorithm to calculate the 'energy' of a signal. *Proc. IEEE ICASSP-90*, 1990.

[17] P. Maragos, T. F. Quatieri, and J. F. Kaiser. Speech nonlinearities, modulations and energy operators. *Proc. IEEE ICASSP-91*, May 1991.