Non Linear Speech Features for the Objective Detection of Discontinuities in Concatenative Speech Synthesis

Yannis Pantazis and Yannis Stylianou

University of Crete, Computer Science Department, Heraklion Crete, Greece, 71110
{pantazis, yannis}@csd.uoc.gr

Abstract. An objective distance measure which is able to predict audible discontinuities in concatenative speech synthesis systems is very important. Previous results showed that linear approaches are not very effective to detect audible discontinuities. The best result was obtained by using the Kullback-Leibler distance on power spectra with the rate of 37%. In this paper, we present two nonlinear approaches for the detection of discontinuities. The first method is based on a nonlinear harmonic model for speech while the second method is based on the demodulation of speech in an amplitude and a frequency component using the Teager energy operator. Results show that detection rate can exceed 70%, which is an improvement of about 95% over previous published results.

1 Introduction

Many modern speech synthesis systems based on non-uniform unit concatenation are quite popular due to their ability to procude high quality and natural-sounding synthetic speech signals [1], [2], [3], [4]. These systems make use of large databases containing many instances of each speech unit (e.g, diphones). In an attempt to minimize audible discontinuities at the concatenation point, these systems try to select the optimum unit from the database. This is done by assigning a target and a concatenation cost to each candidate unit. Target cost, which express the closeness between the context of the target and that of the candidate unit, is evaluated as a weighted sum of differences between prosodic and phonetic parameters. Concatenation cost, which refers to how well adjacent units can be joined, is calculated as a weighted sum of differences between F0, mismatches in spectral features, energy, etc. Total cost is the sum of target and concatenation cost. Optimum unit selection is then achieved by a Viterbi search for the lowest total cost path through the lattice of candidate units. Among these two costs, the concatenation cost is the most important for the selection of two successive acoustic units. Recent studies attempted to specify which concatenation distance measures are able to predict audible discontinuities. Thus, units that are identified to produce audible discontinuities will have less chances of being selected.

Concentrating on concatenation cost, researchers put a lot of effort looking for an objective distance measure which highly correlates with human perception of

discontinuity at unit concatenation point. Klabbers and Veldhuis [5] found that the best predictor of discontinuities was the Kullback-Leibler distance on LPC power spectra. Wouters and Macon [6] found that the Euclidean distance on mel-scale LPC-based cepstral coefficients performed well. 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. Since these studies were conducted on different databases, it is not possible to make direct comparisons between features and methods that were used and draw useful conclusions from them. Despite this fact, most of them showed that Kullback-Leibler distance was on the right track. However, the scores were not very high.

In this paper, we introduce two new sets of features for detecting discontinuities and a new discrimination function in order to increase detection rate. The first set of features are obtained by modeling the speech signal as a sum of harmonics with time varying complex amplitude [9]. The second set of features is based on a technique which tries to decompose speech signals into AM and FM components [10]. Speech signals pass through a filterbank which covers the most important frequencies of the speech spectrum, and then an algorithm referred to as DESA is applied for the separation of the AM and FM component. In contrast with the previous reported studies, we work with vectors instead of scalars which make the discrimination procedure more intricate. We further suggest using Fisher's linear discriminant as a discrimination function.

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 database used and how we construct it. 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 Set of Features

In previous work, speech signals were considered stationary 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 speech resonances can change rapidly within few —even a single— speech periods [11], [12]. Therefore, in an attempt to incorporate dynamic information in the decision whether or not there is an audible discontinuity, we introduce two techniques for the extraction of nonlinear as well as of linear features. Linear features are estimated for comparison purposes only.

2.1 Nonlinear Harmonic Model

The first technique for analysing speech signals is through a nonlinear harmonic model [9]. The model assumes the speech signal to be composed as 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] can take one of the following forms:

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

$$A_k[n] = a_k(n_i) + (n - n_i) b_k(n_i) \tag{3}$$

where $a_k(n_i)$ and $b_k(n_i)$ are assumed to be complex numbers which denote the amplitude of the kth harmonic and the first derivative(slope) respectively. The first method, which leads to a linear harmonic model, is only evaluated for comparison purposes.

The size of analysis window is two pitch periods and it is centered at the concatenation point. It is important to make the analysis at the concatenation point because in our decisions, as explained above, we use dynamic information which may change rapidly within few pitch periods. Therefore, n_i denotes the time instant of the concatenation point. First, the current fundamental frequency, $f_0(n_i)$, is evaluated from the autocorrelation function of the speech signal around the concatenation point. Then, in order to consider the whole spectrum, the number of harmonics, $L(n_i)$, is computed by $L(n_i) = \lfloor \frac{f_s}{2f_0(n_i)} \rfloor$ where f_s denotes the sampling frequency and $\lfloor \rfloor$ denotes the floor operator.

The unknown complex amplitudes (eq. (2) & eq. (3)) are estimated by minimizing a weighted time-domain least-squares criterion with respect to $a_k(n_i)$ or to $a_k(n_i)$ and $b_k(n_i)$,

$$\epsilon = \sum_{n=n_i-T_0}^{n=n_i+T_0} w^2[n](s[n] - h[n])^2 \tag{4}$$

where s[n] denotes the original speech signal, h[n] denotes the harmonic signal to estimate, w[n] denotes the weighted window (which is typically a Hanning window) and T_0 denotes the local fundamental period ($f_s/f_0(n_i)$), in samples. Using Simple Harmonic Model(SHM, eq. (2)) a mean squared error in the order of 5dB is achieved, while using Harmonic Model With Slopes(HMWS, eq. (3)) mean squared error is about 25dB. Obviously, the nonlinear approach models speech signals better.

2.2 AM-FM Decomposition

Teager [11], [12], in his work on nonlinear modeling of speech production, used the nonlinear operator

$$\Psi\{x[n]\} = x^2[n] - x[n-1]x[n+1] \tag{5}$$

on speech signals $x[n]$. This operator, also known as Teager energy operator, was used by Maragos et al. [10] for the separation of amplitude from frequency modulations of a AM-FM signal. The core of the Discrete Energy Separation Algorithm(DESA) are the following equations:

$$G[n] = 1 - \frac{\Psi\{y[n]\} + \Psi\{y[n+1]\}}{4\Psi\{x[n]\}} \tag{6}$$

$$\Omega[n] \approx \arccos(G[n]) \tag{7}$$

$$|a[n]| \approx \sqrt{\frac{\Psi\{x[n]\}}{1 - G^2[n]}} \tag{8}$$

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 [13]. The extraction of a single resonance is done by bandpass filtering the speech signal with a Gabor filter with impulse response defined by

$$h_G[n] = \exp(-b^2 n^2)\cos(\Omega_c n) \tag{9}$$

where b controls the bandwidth of the filter and Ω_c is the central frequency of the resonance.

In our case, we decided to construct a filterbank of twenty Gabor filters. In our filter design the value of b was selected to be 250, hence the bandwidth of each filter was approximately 425Hz. Mel-frequencies were the central frequencies of the filterbank. This choice was motivated by the importance of these frequencies (as this has repeatedly shown in speech literature) in the perception of speech sounds. The size of analysis window was 300 samples (approximately 20msec) centered at the concatenation point.

3 Discrimination Functions & Features

Up to now, research on predicting audible discontinuities in concatenative speech synthesis was concentrated on finding the right features and on finding a distance measure to be applied on these features. 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 audible discontinuous signals and another for signals that were detected to be continuous and try to separate the two classes with statistical methods. 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 l_p Norms

A well known category of norms are l_p norms, where p can take real positive values. They are defined by

$$l_p\{\mathbf{x}\} = ||\mathbf{x}||_{\mathrm{p}} = (\sum_{\mathrm{i=1}}^{\mathrm{d}} |\mathrm{x_i}|^{\mathrm{p}})^{1/\mathrm{p}} \tag{10}$$

where $\mathbf{x} = [x_1, x_2, ..., x_d]^T$, donotes a real or a complex valued vector. For $p = 2$, (l_2) the well known Euclidean distance is obtained, while for $p = 1$ (l_1) is the absolute sum of the elements of the vector. Both norms have used for measuring the differences between spectral amplitude features, in previous work [7], [14]. Euclidean distance on mel-scaled LPC had given the best results at [6].

Apart from these well known norms, we suggest $l_{1/2}$ for measuring differences. Despite this norm's not satisfying the triangular inequality, it has other useful mathematical properties. Intuitively, $l_{1/2}$ norm favors smaller differences than larger ones. This property makes $l_{1/2}$ norm attractive for measuring differences between frequency parameters.

3.2 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

$$y = \mathbf{w^T x} \tag{11}$$

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 $\mathbf{w}$ (Fig. 1).

The direction of $\mathbf{w}$ is important for adequate separation and is given by

$$\mathbf{w} = \mathbf{S_W^{-1}}(\mathbf{m_0} - \mathbf{m_1}) \tag{12}$$

where

$$\mathbf{S_W} = \sum_{\mathbf{i=0}}^{\mathbf{1}} \sum_{\mathbf{x}\in D_i} (\mathbf{x} - \mathbf{m_i})(\mathbf{x} - \mathbf{m_i})^{\mathbf{T}} \tag{13}$$

and

$$\mathbf{m_i} = \frac{1}{\mathrm{N_i}} \sum_{\mathbf{x}\in D_i} \mathbf{x}\ , \qquad \mathrm{i} = 0, 1. \tag{14}$$

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

$$FLD\{\mathbf{x}\} = \sum_{i=1}^{d} w_i x_i \tag{15}$$

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). Now, we are able to combine features which are in different scale.

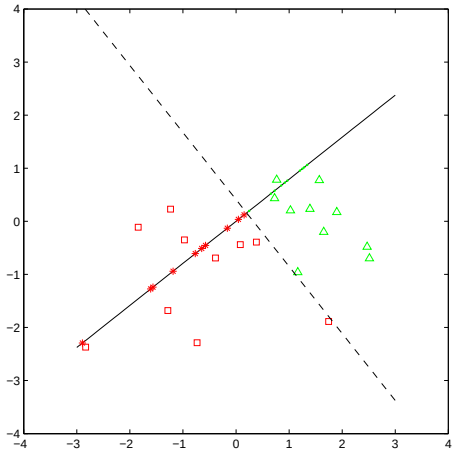

Fig. 1. Example of Fisher's Linear Discriminant

3.3 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}. In our experiments, false alarm was set to 5%. For each measure, y, two probability density functions, $p(y|0)$ and $p(y|1)$ were computed depending on the results from the perceptual test; if the synthetic sentence was perceived as continuous (0), and (1) if it was perceived as discontinuous by the listeners. Then the detection rate for that measure, y, is computed as:

$$P_D(\gamma) = \int_{\gamma}^{\infty} p(y|1)\, dy \tag{16}$$

where γ is defined by:

$$P_{FA}(\gamma) = \int_{\gamma}^{\infty} p(y|0)\, dy \tag{17}$$

3.4 Features

Synthetic test words, as this will be explained in the next section, consist of two parts; a left part and a right one. For both parts, features are computed at the concatenation point. 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, the features of the harmonic models are complex numbers, hence the absolute of their complex difference is considered the same as Euclidean distance between two points on the complex plane. For the second set of parameters, the AM features are defined as the l_1 norm between the AM components estimated for the left and right part for each filter of the filterbank. Similarly, the FM features are estimated as the $l_{1/2}$ norm of the corresponding FM components.

4 Listening Test

Database used for our research was consisted of 2016 monosyllabic words which were generated by concatenative synthesis using an acoustic inventory of recordings from a native American female speaker. The sampling frequency of these recording was 16kHz. The context of the inventory contained 336 monosyllabic test words that constitute the Modified Rhyme Test(MRT)[15]. Synthetic words were obtained by simple concatenation of raw waveforms using each time two halves of original words. The concatenation point was approximately obtained in the middle of the vowel. In order to avoid linear phase mismatches between the concatenated parts, a cross correlation function was used. From listening tests we may say that, in general, pitch continuation was preserved. The 336 spoken words were separated into 56 groups of 6 words. Each group had words with same vowel nucleus but different initial or final consonant(s). Therefore, for each group 36 synthetic words (test stimuli) were constructed (all possible combinations of the 6 recorded words). These 36 synthetic words constitute a subtest. Every subtest contained 6 "synthesized" words which actually were human spoken words and we used them for validation purposes.

The listening task was conducted in a quiet office room using headphones. Listeners were presented with a test stimulus along with a decision in order to familiarize themselves with the listening test. After this training period, listeners started to hear the test words followed by a single interval of forced choice (Yes/No) depending on whether or not they had heard a concatenation discontinuity. The number of subtests listened by the participants was 386.

Twelve listeners participated in the perceptual test. Four of them were native Americans while the others were Greeks with satisfactory knowledge of English language. Five of the participants had experience in listening to synthetic speech. As a validation check, we tested how many of the continuous words were considered as discontinuous. A subtest was rejected if more than one continuous word was considered as discontinuous. This way, 62 subtests were rejected from the database while 324 subtests remained.

Finally, two numbers were assigned to each test stimulus. First number counted how many listeners perceived test stimulus discontinuous while second number counted how many listeners perceived test stimulus continuous. A synthetic speech signal was considered discontinuous(or continuous) if the first number was greater(or less) to the second number. Rarely, when a tie occured synthetic signal was considered as discontinuous.

5 Results

In Table 1, detection rate of various measure distances are presented. We remind that the false alarm was set to 5%.

The parameters of the harmonic models are complex numbers and as mentioned before we use as a difference between complex numbers the absolute of the complex difference. In order to keep the size of the measured vectors small while

Distance	Detection Rate (%)
l_1 on a_k of SHM	32.34
l_2 on a_k of SHM	39.77
l_1 on a_k of HMWS	40.83
l_2 on a_k of HMWS	43.92
Fisher on a_k of SHM	45.46
Fisher on a_k of HMWS	44.50
Fisher on a_k & b_k of HMWS	54.63
Fisher on AM	28.86
Fisher on FM	29.92
Fisher on AM & FM	39.29
Fisher on a_k & b_k & AM & FM	70.46

Table 1. Detection Rates

preserving the important information from a speech frame, we have decided to prune the size vector of complex amplitudes to the twenty first frequencies. Indeed, given that the average fundamental frequency of the voice is about 200Hz we cover most of the time the first 4000Hz of a speech frame. We have considered l_1 norm, l_2 norm and Fisher linear discriminant for both harmonic models. Fisher's linear discriminant on a_k & b_k from the nonlinear harmonic model has given the best score(54.63%).

The second feature set composed by features of the AM & FM model performed poorer than harmonic models. However, these results were higher than previous reported work. Detection rate with the use of Fisher's linear discriminant on the FM components performed slightly better than the AM components. A simple combination of these two components has resulted in a higher detection rate(39.29%). Finally, by applying Fisher's linear discriminant on the whole set of features(Harmonic parameters, AM, FM) an impressive detection rate of 70.46% has been obtained.

6 Conclusion and Future Work

This paper introduced two new feature sets for the problem of detecting audible discontinuities in concatenative speech synthesis. The first set of features, which gave the best result, 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 discontinuities were separated from those without audible discontinuities by a hyperplane which was determined by Fisher's linear discriminant.

A remarkable detection rate(compared to previous published results) was obtained when the above features were combined. However, we expect that better results can be obtained if we use more sophisticated discrimination functions. Moreover, the number of parameters used in this experiment is quite

large. Therefore, data reduction is necessary for a feasible implementation of the suggested approach in the concatenative speech synthesis systems. These two observations draw the line of our future research work.

References

1. A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using large speech database. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 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. 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.
6. 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.
7. Y. Stylianou and A. Syrdal. Perceptual and objective detection of discontinuities in concatenative speech synthesis. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 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. 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.
10. P. Maragos J. Kaiser and T. Quatieri. On separating amplitude from frequency modulations using energy operators. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar 1992.
11. H. M. Teager. Some observations on oral air flow during phonation. *IEEE Trans. Acoust., Speech, Signal Processing*, Oct 1980.
12. 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.
13. P. Maragos T. F. Quatieri and J. F. Kaiser. Speech nonlinearities, modulations and energy operators. *Proc. IEEE ICASSP-91*, May 1991.
14. J. Vepa S. King and P. Taylor. Objective distance measures for spectal discontinuities in concatenative speech synthesis. *ICSLP 2002*, pages 2605–2608, 2002.
15. A. S. House C. E. Williams M.H. L. Hecker and K. D. Kryter. Phycoacoustic speech test: A modified rhyme test. *Tech. Doc. Rept. ESD-TDR-63-403*, Jun 1963.