On the Detection of Discontinuities in Concatenative Speech Synthesis

Yannis Pantazis and Yannis Stylianou

University of Crete, Computer Science Department, Heraklion Crete, Greece, 71409
{pantazis, yannis}@csd.uoc.gr

Abstract. Last decade considerable work has been done in finding an objective distance measure which is able to predict audible discontinuities in concatenative speech synthesis. Speech segments in concatenative synthesis are extracted from disjoint phonetic contexts and discontinuities in spectral shape and phase mismatches tend to occur at unit boundaries. Many feature sets—most of them of spectral nature—and distances were tested. However there were significant discrepancies among the results. In this paper, we tested most of the distances that were proposed using the same listening experiment. Best score were given by AM&FM decomposition of the speech signal using Fisher's linear discriminant.

1 Introduction

Modern Text-to-Speech(TTS) systems are based on concatenated segments of speech units selected from a large inventory [1] [2] [3] [4]. Different instances of each speech segment (or unit) are occurred in the inventory with various prosodic and spectral characteristics. Selection of the appropriate speech units results in high-quality and natural-sounding synthesized speech. In order to select the best units, a combination of two costs is attributed to each candidate unit. The first cost, called target cost, expresses the closeness between the context of the target and the candidate unit [3]. The other cost, called join or concatenation cost, describes how well speech units are concatenated.

Segment mismatches may be caused by various sources such as discrepancies in fundamental frequencies, different levels of loudness (energy of the segments), or variability in spectral contents. The two first, which are of prosodic nature, can be easily adjusted with little degradation in naturalness [5] while spectral mismatches, which are caused by coarticulation phenomena, cannot be changed. Unit selection tries to avoid spectral mismatches by selecting appropriate segments which minimize the concatenation cost. On the other hand, the solution of smoothing usually results in deterioration of the naturalness of the final synthetic speech. Therefore, it is necessary to find an objective spectral distance measure that is able to predict these spectral mismatches. Then, such an objective measure should be the major part of the concatenation cost.

Y. Stylianou, M. Faundez-Zanuy, A. Esposito (Eds.): WNSP 2005, LNCS 4391, pp. 89–100, 2007.

Concatenation cost is usually computed as a distance on a feature vector which is extracted from speech segments [6] (Fig. 1). Recently, a lot of research work has been developed for addressing this problem. However no definite conclusion can be made from these studies since the results were reported on different databases, and conclusions varied. Moreover, each study has conducted each own listening test (i.e. phoneme dependent/independent analysis, with or without signal processing modifications) and, with different setups (i.e. diphone/unit selection synthesizers). This, dramatically influences the quality of the opinions of the perceptual tests. Also, because of the limited duration of the acoustic stimuli (i.e. 100ms) presented to listeners, they usually argued that the assessment of a synthetic segment was difficult. Furthermore, the number of listeners participating in each test is rather limited and thus safe conclusions from only one listening test can not be extracted.

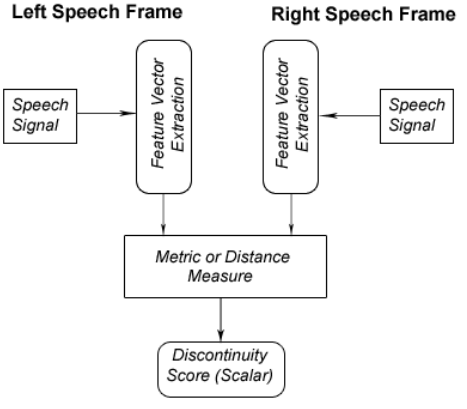

Fig. 1. Flow diagram for measuring the discontinuity of two successive speech units

In this paper we attempted to make a comparison of all these results under a common "space", i.e. compare the methods proposed in a database previously used for the same research purpose. Various spectral features coming from speech coding, speech recognition, speech analysis and synthesis were tested. Distances such as absolute difference metric, Euclidean distance, Kullback-Leibler divergence as well as statistical methods were used for the evaluation of perceived discontinuity.

The paper is organized as follows. In Section 2, the different methods of psychoacoustic experiments are presented while Section 3 presents the various speech features that were used in the current study. Section 4 describes the different distance measures used in the evaluation of discontinuities. Section 5 describes how speech naturalness is improved using these discontinuity measures, while in Section 6 the database where the evaluation has been performed is briefly presented. Finally, in Section 7 results and major conclusions are presented.

2 Review of Perceptual Listening Experiments

Since the purpose of unit selection is to locate segments (units) that will make the synthetic speech to sound natural, much effort has been devoted to finding the relation between objective distance measures and perceptual impressions. Searching for an objective distance measure that is able to predict perceptual discontinuities or is able to measure the variations of allophones, a subjective measure need to be obtained. For this purpose listeners are asked to decide for the existence of discontinuity or to judge speech quality, which may be evaluated for intelligibility, naturalness, voice pleasantness, liveness, friendliness, etc. Because of the expected variability in the human responses Mean Opinion Score (MOS) is usually used to determine the quality of the synthetic speech.

2.1 Perceptual Evaluation of Discontinuity

One approach for the evaluation of perceptual tests is to generate monosyllabic words with a change point at the middle of the vowel [7]. Every word pair in the perceptual test consists of a reference word and a modified version of this word. Instead of monosyllabic words sentences can be also used as in [8]. Listeners have to assess how close these pairs are in a five-point scale. Then correlations are computed between perceptual test and objective distance measures. A variant of this method is to synthesize sentences (or words) with different objective distance measures and ask listeners which sentence is more sonorant [9].

Another approach for the evaluation of objective distance is to construct a concatenation and ask the listeners whether or not a discontinuity is perceived [10], [11]. A less rigid task for the listener was to rate the discontinuity at the concatenation in a five-point scale [12], [13].

2.2 Test Stimuli

An issue which also determines the quality of the perceptual experiment is the contents and the duration of the stimuli. The contents of the stimuli varies from few vowels [7], [10] and diphthongs [13] to the 336 monosyllabic test words that constitute the Modified Rhyme Test [11]. The vowels used in these experiments, were selected in such a way that they corresponded to distinct tongue positions. Few studies have used consonants in their stimuli [9]. Duration also varies from few milliseconds [10], [12] to monosyllabic words [7], [11] and even entire sentences [13], [9].

3 Spectral Feature Representations

Due to the spectral nature of the problem, many spectral feature representations were tested.

3.1 Well Known Feature Sets

FFT-based spectrum (D1) as well as LPC-based spectrum (D2) were tested by many researchers. Another common feature representation of a speech magnitude spectrum is that of Line Spectral Frequencies (LSF) [14] (D3). Depending on the sampling frequency of the speech signal, a few number (i.e. 18–20) of LSFs are usually extracted from the signals. LSFs encode speech spectral information efficiently and provide good performance both in speech coding and speech recognition.

Borrowed from speech recognition systems [15], Mel-scaled Frequency Cepstral Coefficients (MFCCs) is a feature representation that has been extensively used for the detection of audible discontinuities [10] [11] [13]. Like LSFs, the number of MFCCs extracted from the speech signals, depends on the sampling frequency. The dominance of MFCCs in speech recognition as well as in speaker identification/verification systems stems from their ability to represent the amplitude spectrum in a compact form. They may be computed using two different methods; FFT spectrum (D4) and LPC spectrum (D5).

3.2 Less Common Features

3.2.1 Multiple Centroid Analysis

A spectral feature set referred as Multiple Centroid Analysis (MCA) (D6) was introduced in [13] for the prediction of discontinuities. MCA is an alternative to formant estimation techniques. If the spectral distribution within a partition of the spectrum contains a single formant then the centroid and associated variance, represent the formant frequency and bandwidth (Fig. 2). In [13] four centroid and the corresponding bandwidths were extracted from the speech signals.

The evaluation of the centroid was done by minimizing the "error" quantity

$$e(c_i, d_i) = \sum_{i=1}^{N} \sum_{k=c_{i-1}}^{c_i} P[k](k - d_i)^2$$

where $P[k]$ is the power spectrum, c_i represents the bounds (bandwidth) and d_i denotes the centers of the formants. N determines the number of centroid, which also depends on the sampling frequency. For example, if the sampling frequency is equal to $16kHz$, four centroid are evaluated [13].

3.2.2 Bispectrum

Speech features obtained by linear prediction analysis as well as by Fourier analysis are determined from the amplitude or power spectrum. Thus, the phase information of the speech signal is neglected. However, phase information has been proven to play an important role in speech naturalness and signal quality in general. Furthermore, the higher order information is ignored since the power spectrum is only determined by second order statistics. If speech were a Gaussian

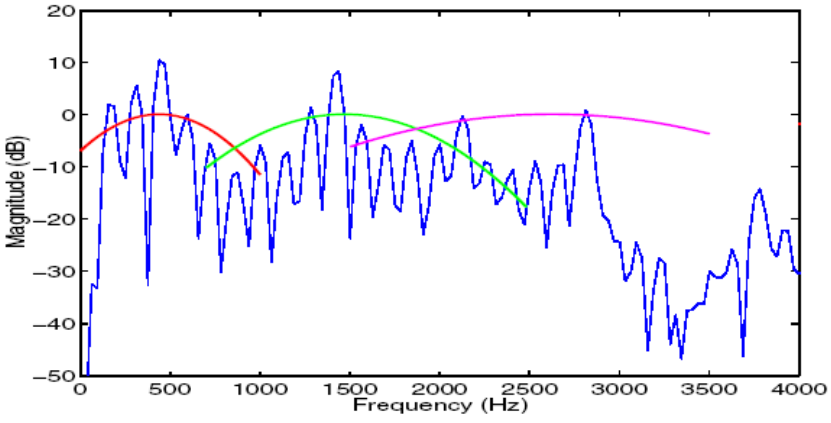

Fig. 2. Centers of gravity (after Vepa and King)

process, then the second order statistics would suffice for a complete description. However, evidence appears to indicate that in general, speech is non-Gaussian. To take into account phase information as well as higher order statistics bispectrum as well as Wigner-Ville transform and modified Mellin transform were tested by Chen et al. [8]. In this paper bispectrum (D7) was also tested, since it has been shown [8] that it provides high correlation scores.

Bispectrum is defined as a 2-D Fourier transform of 2-lag autocorrelation function.

$$S_{3x}(f_1, f_2) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} C_{3x}(k, l) e^{-j2\pi f_1 k} e^{-j2\pi f_1 l} \tag{1}$$

where

$$C_{3x}(k, l) = \sum_{n=-\infty}^{\infty} x^*[n] x[n+k] x[n+l] \tag{2}$$

is the 2-lag autocorrelation function

3.2.3 Nonlinear Approaches

Another drawback of linear prediction analysis and Fourier analysis is that speech signals are considered stationary around the concatenation point. Hence, the techniques used for the extraction of the feature set do not take into account any dynamic information of the speech signal. But experimental work provided evidence that speech resonances can change rapidly within a few—even a single—speech periods [16], [17]. Therefore, in an attempt to incorporate dynamic information in the decision whether or not there is an audible discontinuity, two techniques have been introduced for the extraction of nonlinear features [18].

a. Time-Varying Harmonic Model
The first set of features are obtained by modeling the speech signal as a sum of harmonics with time varying complex amplitude (D8). This results in representing speech signal by a nonlinear harmonic model [19]. The model assumes the speech signal to be composed by a periodic signal, h[n], which is designated as a sum of harmonically related sinusoids

$$h[n] = \sum_{k=-L(n_i)}^{L(n_i)} A_k[n] e^{j2\pi k f_0(n_i)(n-n_i)} \tag{3}$$

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

$$A_k[n] = a_k(n_i) + (n - n_i) b_k(n_i) \tag{4}$$

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.

b. AM&FM Decomposition
The second set of features is based on a technique which tries to decompose speech signals into Amplitude Modulated (AM) and Frequency Modulated (FM) components (D9). Teager [16], [17], in his work on nonlinear modeling of speech production, has used the nonlinear operator known as Teager-Kaiser energy operator:

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

on speech signal, $x[n]$. Based on this operator, Maragos et al. [20] have developed the Discrete Energy Separation Algorithm(DESA) for separating an AM-FM modulated signal into its components. An AM-FM modulated signal has the form

$$x[n] = a[n] cos(\Omega[n])$$

where $\Omega[n]$ is the instantaneous frequency and $a[n]$ is the instantaneous amplitude.

3.2.4 Phonetic Features

Prosodic and phonetic features can be used for the evaluation of concatenation cost. This is admissible since different phonetic and/or prosodic contents affect the realization of neighbouring phones —coarticulation phenomena. Phonetic features found to be more efficient than acoustic measures in predicting audible discontinuities [21] [22]. For this reason target cost may be more important since target cost is computed as a weighted sum of subcosts of prosodic and phonetic nature. However, using only the target cost, someone cannot eliminate concatenation discontinuities nor can measure the closeness of two successive speech segments.

4 Distance Measures

After the computation of features at the concatenated segments, the closeness of them should be somehow determined. As measures someone can use metrics, similarity measures and discriminant functions. Here, the following distance measures were tested.

(a) l_1 or absolute difference
(b) l_2 or Euclidean distance
(c) Kullback-Leibler divergence
(d) Mahalanobis Distance
(e) Fisher's linear discriminant
(f) Linear regression

Absolute and Euclidean distance are metrics that belong to the same family. Their difference rely on the fact that Euclidean distance amplifies more the difference of specific parameters of the feature vector than absolute distance.

Kullback-Leibler (KL) divergence as well as Mahalanobis distance come from statistics. Mahalanobis distance is similar to Euclidean with each parameter of the feature vector being divided by its variance. KL divergence is used to measure the distance between two probability distributions.

A symmetric version of KL divergence was used to measure the distance between two spectral envelopes and is given by,

$$D_{KL}(P,Q) = \int (P(\omega) - Q(\omega))\ log\left(\frac{P(\omega)}{Q(\omega)}\right) d\omega \tag{6}$$

4.1 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

$$y = \mathbf{w^T x} \tag{7}$$

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. 3).

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

$$FLD\{\mathbf{x}\} = \sum_{i=1}^{d} w_i x_i \tag{8}$$

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, features which are in different scale can now be combined or compared.

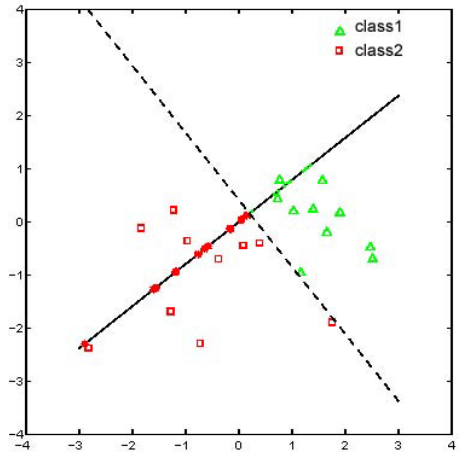

Fig. 3. Example of Fisher's Linear Discriminant between two classes

4.2 Linear Regression

Linear regression fits an input feature set to the observations (or output) using a least-squares criterion. In our case input vectors are the various feature representations and output is the Mean Opinion Score of listeners. Linear regression is similar to Fisher linear discriminant since both methods are linear and optimal for normal distributions. However, their parameters are estimated by different ways.

5 Improve Speech Naturalness

When the inventory is small there will be cases where the selected units are not matching very well. Moreover, the objective distance measures used in modern synthesis systems do not correlate very well with human perception, thus spectral mismatches may occur. Therefore, smoothing at the concatenation points is necessary to lower the mismatch effect.

Wouters and Macon [23] reduced concatenation mismatches by combining spectral information represented by LSFs from two sequences of speech units selected in parallel. The first sequence defined the initial spectral trajectories for a target utterance. Then, this sequence was modified by the second sequence which defined the desired transitions between concatenation units. Perceptual experiments showed that considerable amount of concatenation artifacts were removed.

Klabbers and Veldhuis [24] extended the diphone inventory with context-sensitive diphones. Using their best predictor which was based on spectral features (MFCC), they have clustered the contexts obtaining new recordings. To evaluate the improvements they have conducted further experiments and they have found that the added diphones significantly reduce the amount of audible discontinuities.

In an other study of Vepa and King [25], linear dynamical model (Kalman filter) on LSF trajectories has been used for the computation of join cost in unit selection speech synthesis. The model, after training, could be used to measure how well concatenated speech segments join together. The objective join cost is based on the error between model prediction and actual observations. Linear dynamical model was used also for smoothing the LSF coefficients reducing the audible discontinuities. An advantage of this method is that the degree and extent of the smoothing is controlled by the model parameters which are learned from natural speech.

6 Database and Listening Experiment

In this section we briefly present the database used for comparing all the previously reported methods and features as well as the listening experiment that was conducted. A more detailed description can be found in Klabbers et al. [24]. It is worth to note that since the same database has already been used on the same task, useful conclusions may be reached.

Five subjects with backgrounds in psycho-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 nonsense words of the form $C@CVC@$ (where C =consonant, V =vowel$\in$ /a/, /i/ and /u/ and @ = schwa). The recordings were made of a semiprofessional female speaker. Speech signals have been sampled at $16kHz$.

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. Listeners' 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.

7 Results

7.1 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 probability density functions, $p(y|0)$ and $p(y|1)$ were computed depending on the results from the perceptual test: (0) if the synthetic sentence was perceived as continuous and (1) if it was

perceived as discontinuous by the listeners. Then the detection rate for that measure, y, is computed as:

$$P_D(\gamma) = \int_{\gamma}^{\infty} p(y|1)\, dy \tag{9}$$

where γ is defined by:

$$P_{FA}(\gamma) = \int_{\gamma}^{\infty} p(y|0)\, dy = 0.05 \tag{10}$$

which means that the false alarm rate was set to 5%.

7.2 Results and Discussion

In Table 1, detection rate of various measure distances are presented. For the statistical methods such as Fisher linear discriminant and linear regression, the training was done on the 80% of the database, while the testing was done on the remaining 20% of the database. Note also that the evaluation is independent of the phonemes of the database while most of previous studies were phoneme specific. Phoneme specific approaches [7] [24] [8] provide better results compared to phoneme independent approaches [11]. 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.

Table 1. Detection Rates. False Alarm was set at 5%.

Distance	Detection Rate (%)	Distance	Detection Rate (%)
D1a	10.31	D1b	19.66
D1c	17.27	D2a	15.35
D2b	20.14	D2c	23.50
D3a	17.75	D3b	17.27
D3d	6.24	D3e	38.13
D3f	37.26	D4a	33.33
D4b	36.93	D4d	28.54
D4e	40.53	D4f	39.61
D5a	39.33	D5b	37.65
D5d	27.58	D5e	41.01
D5f	40.78	D6a	10.55
D6b	9.83	D6d	10.07
D6e	25.42	D6f	24.90
D7a	12.04	D7b	19.24
D8e	46.52	D8f	45.50
D9e	49.40	D9f	47.83

In the table, the feature sets are represented with numbers (D1, D2, ...), while the letters (a, b, ...) following the feature set correspond to the distance. For example, D3d means that LSF coefficients have been used along with the Mahalanobis distance. It is obvious from the table that none speech representation passed 50% of detection rate. Spectrum evaluated from FFT (D1), from LPC coefficients (D2) and Bispectrum (D7) gave small detection rate. LSFs and MFCCs combined with Fisher's linear discriminant performed well. Same conclusion can be made for the nonlinear harmonic model and AM&FM decomposition The latter gave the best detection rate 49.40%. Linear regression gave detection rates close to Fisher's linear discriminant as it was expected. These results show clearly that a lot of works remains to be done despite the considerable effort of many researchers on searching an optimal distance and feature representation.

From the above it is obvious that using a weighted distance the detection rates are improved independently of the features. This is explained by the fact that weights are trained from the same database. Moreover these data-driven weights can boost some particulate parameters of the feature vector and eliminate some others.

References

1. Robert E. Donovan. *Trainable Speech Synthesis.* PhD thesis, Cambridge University, Engineering Department, 1996.
2. 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.
3. 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.
4. 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.
5. Thierry Dutoit. *An Introduction to Text-to-Speech Synthesis.* Kluwer Academic Publishers, 1997.
6. T. R. Barnwell S. R. Quackenbush and M. A. Clements. *Objective Measures of Speech Quality.* Prentice Hall, 1988.
7. 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.
8. J.-D. Chen and N. Campbell. Objective distance measures for assessing concatenative speech synthesis. *EuroSpeech99*, pages 611–614, 1999.
9. Jerome R. Bellegarda. A novel discontinuity metric for unit selection text-to-speech synthesis. *5th ISCA Speech Synthesis Worksop*, pages 133–138, 2004.
10. 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.
11. Y. Stylianou and A. Syrdal. Perceptual and objective detection of discontinuities in concatenative speech synthesis. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001.

12. 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.
13. J. Vepa S. King and P. Taylor. Objective distance measures for spectal discontinuities in concatenative speech synthesis. *ICSLP 2002*, pages 2605–2608, 2002.
14. F. K. Soong and B. H. Juang. Line spectrum pairs and speech data compression. *ICCASP*, pages 1.10.1–1.10.4, 1984.
15. L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.
16. H. M. Teager. Some observations on oral air flow during phonation. *IEEE Trans. Acoust., Speech, Signal Processing*, Oct 1980.
17. 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.
18. Y. Pantazis Y. Stylainou and E. Klabbers. Discontinuity detection in concatenated speech synthesis based on nonlinear analysis. *InterSpeech2005*, pages 2817–2820, 2005.
19. Yannis Stylianou. *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, 1996.
20. 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.
21. H. Kawai and M. Tsuzaki. Acoustic measures vs. phonetic measures as predictors of audible discontinuity in concatenative synthesis. *ICSLP*, 2002.
22. A. K. Syrdal and A. D. Conkie. Data-driven perceptually based join cost. *5th ISCA Speech Synthesis Workshop*, pages 49–54, 2004.
23. J. Wouters and M. W. Macon. Unit fusion for concatenative speech synthesis. *ICSLP*, Oct 2000.
24. E. Klabbers and R. Veldhuis. Reducing audible spectral discontinuities. *IEEE Transactions on Speech and Audio Processing*, 9:39–51, Jan 2001.
25. J. Vepa and S. Taylor. Kalman-filter based join cost for unit selection speech synthesis. *Eurospeech*, Sep 2003.