INTERSPEECH 2009 BRIGHTON

A Close Look into the Probablistic Concatenation Model for Corpus-based Speech Synthesis

Shinsuke Sakai[1], *Ranniery Maia*[1], *Hisashi Kawai*[1], *Satoshi Nakamura*[1]

[1] National Institute of Information and Communications Technology, Japan

{shinsuke.sakai,ranniery.maia,hisashi.kawai,satoshi.nakamura}@nict.go.jp

10.21437/Interspeech.2009-254

Abstract

We have proposed a novel probabilistic approach to concatenation modeling for corpus-based speech synthesis, where the goodness of concatenation for a unit is modeled using a conditional Gaussian probability density whose mean is defined as a linear transform of the feature vector from the previous unit. This approach has shown its effectiveness through a subjective listening test. In this paper, we further investigate the characteristics of the proposed method by a objective evaluation and by observing the sequence of concatenation scores across an utterance. We also present the mathematical relationships of the proposed method with other approaches and show that it has a flexible modeling power, having other approaches to concatenation scoring methods as special cases.

Index Terms: speech synthesis, unit selection, join costs

1. Introduction

It is crucial to establish a good concatenation cost for the quality of concatenative speech synthesis and there has been a number of research efforts to find a good measure of concatenation cost [1, 2, 3, 4], in which various spectral feature parameters and distance measures are investigated. There also is a research effort to find optimal mapping functions from distance measures to costs based on perceptual evaluation [5]. In a previous paper, we departed from the traditional view of cost based on "distance" and attempted to take a probabilistic view of concatenation cost where concatenation modeling is done with a probabilistic model that captures how likely it is to observe the spectral shape of the current unit given the spectral shape of the previous unit, using conditional Gaussians [6]. We performed a subjective listening test and confirmed the effectiveness of the proposed method [6].

In this paper, we further investigate the characteristics of the proposed method by an objective evaluation comparing the closeness of the synthetic speech samples to natural speech as measured by the distance of MFCC parameter sequences. We also look at the sequence of concatenation scores across an utterance and see how it is behaving similarly or differently compared to the baseline method. We also present the mathematical relationships of the proposed method with other approaches and show that it has a flexible modeling power, having various other scoring methods as special cases.

In the next section, we summarize the proposed concatenation modeling based on conditional Gaussians. Objective evaluation experiments are reported in the succeeding section, followed by the section investigating the sequences of concatenation scores for the 'correct' unit sequence. We then explore the mathematical relationships of the proposed method with other methods followed by the conclusion.

2. Concatenation modeling using conditional Gaussian

We model the goodness of concatenation in terms of the probability that a spectral shape of a unit, $o(u_i)$), is observed after the previous unit in the phonetic context determined by the input specification S and the current unit position i, through conditional Gaussian density,

$$\begin{aligned} P(o(u_i)|o(u_{i-1}), S, i) &= P(h(u_i)|t(u_{i-1}), S, i) \\ &= \mathcal{N}(h(u_i)|B\,t(u_{i-1}) + b,\ \Sigma), \end{aligned}$$

where d-dimensional vector $h(u_i)$ represents the average spectrum of an initial part (or *head*) of the unit u_i, and the d-vector $t(u_i)$ represents that of a final part (or *tail*) of the unit u_{i-1}. B is a $d \times d$ regression matrix with the j-th row representing a regression coefficients for the j-th component of $h(u_i)$, and b is a d-dimensional vector of intercepts, and Σ is a $d \times d$ covariance matrix. B, b, and Σ are determined by the phone identities of the units $u_i$and u_{i-1}.

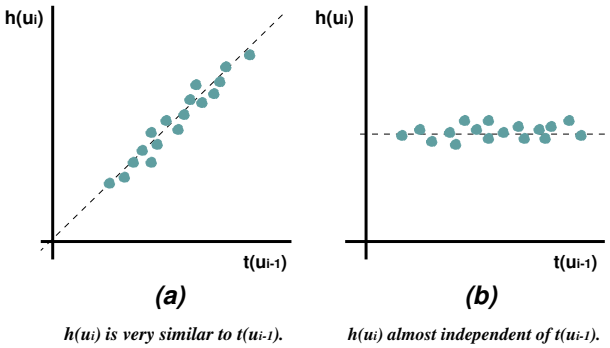

Figure 1: One-dimensional schematic diagram representing the relationship between $h(u_i)$ and $t(u_{i-1})$ in two extreme cases.

If we think about an extreme situation where spectral shapes are very similar across the concatenation boundary, the regression matrix B is considered to be close to identity matrix and the constant vector b is close to zero, as shown in Fig. 1(a). On the other hand, if the head spectrum is almost independent of the tail of the preceding unit, the regression matrix B is considered to be close to zero and b will be the significant constituent of the mean vector. In general cases between the two extremes, B and b are considered to have some meaningful values that represent u_i's characteristics that is dependent on u_{i-1} in some aspects and independent of it in some other aspects.

2.1. ML estimation of conditional Gaussian model parameters

The maximum likelihood (ML) estimate of the model parameters, B and b, from the training data $\mathcal{D} = \{(t_1, h_1), ..., (t_N, h_N)\}$ is derived as a solution to a simple convex optimization problem. The training data $\mathcal{D} = \{(t_1, h_1), ..., (t_N, h_N)\}$ for a conditional Gaussian model for a particular class of unit boundary (a phone pair, in the current implementation with phone-sized units) consists of all the pairs (t_k, h_k) of tail and head spectral feature vectors available from the corpus for that class of unit boundary.

By defining a $d \times (d+1)$ matrix A and a $(d+1)$-vector s_i, such that,

$$A = \left[\ b \mid B\ \right], \quad and \quad s_k = \left[\begin{array}{c} 1 \\ t_k \end{array}\right], \tag{1}$$

we see a relationship $B\,t_k + b = As_k$, and we obtain the estimates of B and b from the estimate of A. The log likelihood $\mathcal{L}$ of the training data $\mathcal{D}$ is, therefore,

$$\begin{aligned} \mathcal{L}(A, \Sigma; \mathcal{D}) &\triangleq \log \prod_{k=1}^{N} \mathcal{N}(h_k | As_k, \Sigma) \\ &= -\frac{dN}{2}\log 2\pi - \frac{N}{2}\log|\Sigma| \\ &\quad -\frac{1}{2}\sum_{k=1}^{N}(h_k - As_k)^T\Sigma^{-1}(h_k - As_k). \end{aligned} \tag{2}$$

Taking the partial derivative of $\mathcal{L}$ with regard to A,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial A} &= -\frac{1}{2}\Sigma_{k=1}^{N}\{-(\Sigma^{-1} + \Sigma^{-1^T})(h_k - As_k)s_k^T\} \\ &= \Sigma^{-1}\sum_{k=1}^{N}(h_k - As_k)s_k^T. \end{aligned} \tag{3}$$

By setting the partial derivative to zero, we obtain the ML estimate of A,

$$\hat{A} = (\sum h_k s_k^T)(\sum s_k s_k^T)^{-1}. \tag{4}$$

The covariance matrix Σ can be estimated as the sample covariance around the conditional mean $\hat{A}\,s_k$, and it reduces to

$$\hat{\Sigma} = \frac{1}{N}\sum_{k=1}^{N} h_k {h_k}^T - \hat{A}\frac{1}{N}\sum_{k=1}^{N} s_k {h_k}^T. \tag{5}$$

Since the number of different combinations of tail phones and head phones is roughly the square of the number of phones, we may not have enough training data for some combinations of tail and head phones, and this sparse data situation can vary depending on how large the available training data is. In order to achieve robust training of conditional Gaussians (CGs), we tie the model parameters using phonetic decision-tree clustering. The models are clustered according to the phonetic questions about the tail phones [6].

3. Objective evaluation experiments

In our previous paper [6], we demonstrated the effectiveness of the proposed approach through a subjective listening test using a corpus-based speech synthesizer reported in [7], with Euclidean distance as the baseline, which has been reported to be a good predictor of perceived discontinuity when measured on Mel-cepstral feature parameters [8]. In the test, eight subjects listened to the speech synthesis output from two synthesizers, one of which adopting Euclidean distance and the other with the proposed conditional Gaussian (CG) models for concatenation cost. They were asked to give scores of 1 to 5 to each utterance. The results of the listening test is shown in Table 1. The mean opinion score with the proposed method was significantly higher than the baseline at the 1% level by the paired t-test.

Table 1: 5-level mean opinion scores for the two synthesizers [6].

Euclidean	CG
2.44	2.97

In this paper, we further investigate the effectiveness of the proposed method through the evaluation in an objective way, by comparing the closeness of the synthetic speech to natural speech as measured by the distance of the MFCC parameter sequence. The differences of the lengths of the parameter sequences were absorbed using dynamic time warping [9]. In the closed part of the evaluation, a set of 29 prompt sentences from the corpus for developing the synthesizer [6] was synthesized using the baseline and proposed methods for concatenation modeling. Fig. 2 plots the average distances between synthetic and natural speech for the baseline (Euclidean distance) and the proposed method (conditional Gaussian). Table 2 shows the means and the standard deviations of the MFCC distances for the baseline and the proposed approach. From the figure and the table, we see that the proposed method achieves smaller distance to natural speech. This difference turned out to be statistically significant at the 1% level.

Table 2: Test results for the closed data, i.e. training sentences. 'cg dists' represents the proposed method that employs conditional Gaussian-based concatenation models and 'euc dists' represents the baseline with Euclidean distance. Standard deviation of the dtw distances are shown in the column headed by "s.d."

	number	mean	s.d.
cg dists	29	16.1	2.26
euc dists	29	17.7	0.98

We also performed an open test by synthesizing 50 conversational sentences (categorized as `conv`) used in the Blizzard Challenge 2005 [10]. Fig. 3 plots the average distances between synthetic and natural speech for the baseline (Euclidean distance) and the proposed method (conditional Gaussian) in the open test. Table 3 shows the means and the standard deviations of the MFCC distances for the baseline and the proposed approach. In the open case, the difference between the proposed method and the baseline is smaller, but the proposed method was still significantly closer to the reference natural speech at the 5% level.

By comparing Fig. 2 and Fig. 3, we note that the distances

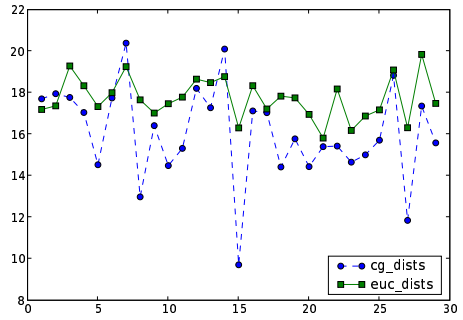

Figure 2: Average MFCC distance between synthetic and natural speech for the baseline (green squares) and the proposed approach (blue dots). The 29 points along the horizontal axis represents the 29 utterances.

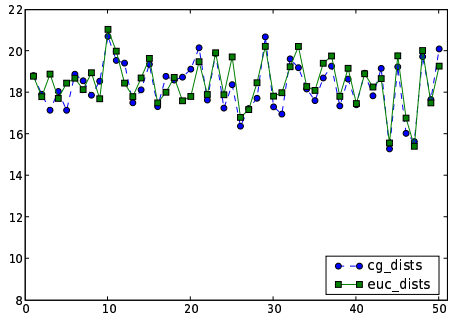

Figure 3: Average MFCC distance between synthetic and natural speech for the baseline (green squares) and the proposed approach (blue dots). The 50 points along the horizontal axis represents the 50 utterances.

to the reference natural speech is smaller for the synthesized speech of closed sentences with both Euclid distance (baseline) and the conditional Gaussian (proposed), due to the fact that the target and concatenation models are trained, as well as the unit database was developed using the data set that includes this reference natural speech. This also confirms that the distance measure used here in the objective evaluation has an expected characteristics.

We also note that the variance of the distances among sentences for Euclidean distance is not very much different between closed data and the test data since this distance measure is not based on a model estimated from the training data.

On the other hand, we note that the distance between the synthesized speech and the reference natural speech gets very small for some sentences (for example, the sentence 15 and the sentence 27 in Fig. 2). A possible reason for this will be that the feature vectors of synthesis units for these sentences happened to be very close to the values (i.e. means of the Gaussian models) predicted by the target and concatenation models.

Table 3: Test results for the open data. 'cg dists' represents the proposed method that employs conditional Gaussian-based concatenation models and 'euc dists' represents the baseline with Euclidean distance. Standard deviation of the dtw distances are shown in the column headed by "s.d."

	number	mean	s.d.
cg dists	50	18.31	1.21
euc dists	50	18.48	1.13

4. A close look at the concatenation scores

In order to investigate the properties of the concatenation scores in more detail, concatenation scores based on conditional Gaussian models and Euclidean distance for a sequence of database units that corresponds to one sentence are plotted in Figure 4. For the ease of visual comparison, squares of the Euclidean distances are also plotted (red dots and solid lines.) This utterance is also part of the training data for target and concatenation models.

Interestingly, by comparing the graphs for conditional Gaussian model and Euclidean distance, wee see that the rough trends are similar between the two concatenation score methods. One explanation for this similarity may be that diagonal components in the transform matrix B are prominent in many of the conditional Gaussian models and they are both influenced by the goodness of the boundaries determined automatically by the forced alignments using a speech recognizer.

On the other hand, we note that there are several points where scores based on Euclidean distance has a big dip whereas scores by conditional Gaussian does not have such a big dip, e.g. transitions at 6 ('-' to 'ay') and 49 ('n' to 's'). A natural interpretation of this would be that it is unlikely to have a small Euclidean distance between phones with very different spectral shapes when the boundary assignment is accurate, whereas it is possible to have a good score with conditional Gaussian models since the conditional mean given by transforming the left unit feature vector can have a shape similar to the right unit feature vector.

5. Relationships with other approaches

The proposed concatenation score between the feature vector from the left unit, t and the feature vector from the right unit h is a log probability given by the conditional Gaussian model and expressed as

$$\begin{aligned}\log \mathcal{N}(h|B\,t+b,\ \Sigma) \\ = \quad & -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| \\ & -\frac{1}{2}(h-(B\,t+b))^T\Sigma^{-1}(h-(B\,t+b)), \end{aligned} \quad (6)$$

where the model parameters, B, b, and Σ depends on the pair of phone identities for the left and right units. The integer d is the dimensionality of the vectors h and t. By comparing this equation with the formulas for other distance measures, we realize that the proposed method has an interesting relationships with other approaches.

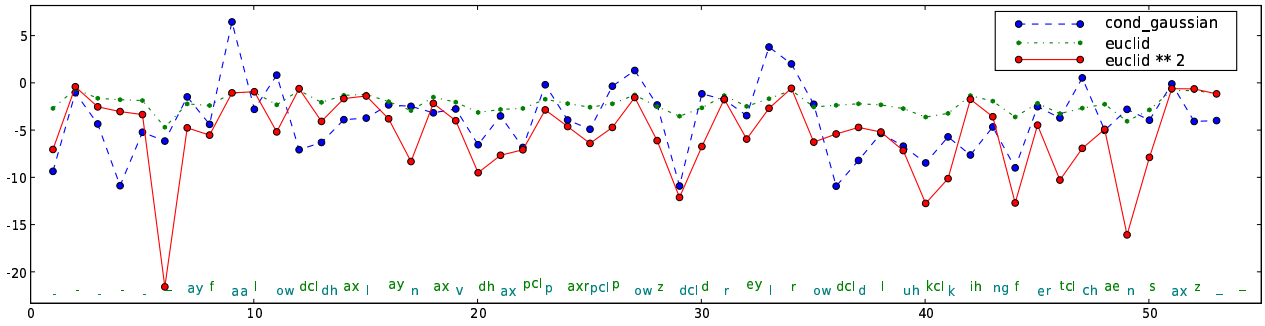

Figure 4: Plot of the concatenation scores for the "true" database units for the sentence *"I followed the line of the proposed railroad, looking for chances."* The blue dots with broken lines represent concatenation scores with conditional Gaussian models, the green dots with dotted lines are concatenation scores that are the negatives of Euclidean distances, and the red dots with solid lines represent the negatives of the squares of Euclidean distances. Scores are for concatenation of the unit at the dot and the one at the next dot.

5.1. Euclidean distance

Euclidean distance is a widely used distance measure and, as mentioned in the previous section, it is reported to be a good predictor of perceived discontinuity [8]. If we set the transformation matrix B to the identity matrix (I), the constant $b = 0$, the covariance matrix Σ also to the identity matrix and neglect the constant terms, we note that the negative of the score given by equation (6) turns out to the square of the Euclidean distance between h and t that can be expressed as

$$D_{euc} = (h - t)^T (h - t). \tag{7}$$

5.2. Donovan's approach

In [3], Donovan proposed a distance measure between the vector e at the end of one segment and the vector s at the start of the next segment. For this purpose, he clustered the pairs of frames across the boundaries using decision tree by asking broad class questions about the preceding and following phonetic identity and the location of the boundary within the phone, and calculated the mean and the covariance matrix within each leaf of the tree. He describes it "a decision-tree-based context-dependent Mahalanobis distance", which is expressed as

$$D^2 = \sum_{i=1}^{n} \left[\frac{e_i - s_i - \mu_i^l}{\sigma_i^l} \right]^2, \tag{8}$$

where n is the dimensionality of the data, μ_i^l is the i-th element of the mean vector in leaf l, σ_i^l is the i-th diagonal element of the covariance matrix for leaf l.

Looking at the equations (6) and (8), we note that (6) becomes equivalent to (8) if we set B to identity matrix and neglect the second term with the determinant of the covariance matrix, also assuming that the elements of the feature vectors are independent to each other. In other words, Donovan's distance measure is similar to the conditional Gaussian-based concatenation model with conditional mean formed by just the addition of the constant b and no transform by the matrix B.

6. Conclusion

In this paper, we presented our attempt of objective evaluation for the concatenation modelling approach based on conditional Gaussian, in which the proposed approach was shown to yield synthetic speech closer to natural speech as measured by distance between MFCC sequences. We also observed the sequence of concatenation scores across an utterance and confirmed that the characteristics of the model is reflected in the behavior of the scores for a "correct" unit sequence. We also presented the mathematical relationships of the proposed method with other approaches and showed that it has a flexible modeling power, having various other scoring methods as special cases.

7. Acknowledgements

The authors are grateful to Prof. Alan Black at Carnegie Mellon University for letting us use a part of test data for Blizzard Challenge 2005.

8. References

[1] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 1, pp. 39–51, 2001.

[2] Y. Stylianou and A. K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proc. ICASSP 2001*, Salt Lake City, USA, 2001.

[3] R. Donovan, "A new distance measure for costing spectral discontinuities in concatenative speech synthesisers," in *Proc. 4th ESCA Tutorial and Research Workshop on Speech Synthesis*, Scotland, Sep. 2001.

[4] J. Vepa and S. King, "Join cost for unit selection speech synthesis," in *Text to Speech Synthesis*, A. Alwan and S. Narayanan, Eds. Prentice Hall, 2004.

[5] T. Toda, H. Kawai, and M. Tsuzaki, "Optimizing sub-cost functions for segment selection based on perceptual evaluations in concatenative speech synthesis," in *Proc. ICASSP 2004*, Montreal, Canada, May 2004, pp. 657–660.

[6] S. Sakai and T. Kawahara, "Decision tree-based training of probabilistic concatenation models for corpus-based speech synthesis," in *Proc. Interspeech 2006*, Pittsburgh, PA, Sep. 2006.

[7] S. Sakai and H. Shu, "A probabilistic approach to unit selection for corpus-based speech synthesis," in *Proc. Interspeech 2005*, Lisbon, Portugal, Sep. 2005, pp. 81–84.

[8] J. Wouters and M. Macon, "A perceptual evaluation of distance measures for concatenative speech synthesis," in *Proc. ICSLP 98*, Sydney, Australia, 1998, pp. 2747–2750.

[9] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall PTR, April 1993.

[10] A. Black and K. Tokuda, "Blizzard challenge – 2005: Evaluating corpus-based speech synthesis on common datasets," in *Proc. Interspeech 2005*, Lisbon, Portugal, 2005, pp. 77–80.