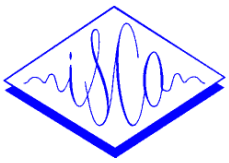


2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

SPEECH SEGMENT NETWORK APPROACH FOR AN OPTIMAL SYNTHESIS UNIT SET

Naoto Iwahashi & Yoshinori Sagisaka

ATR Interpreting Telephony Research Laboratories
2-2 Hikaridai, Seika-cho, Kyoto 619-02, Japan

ABSTRACT

In this paper, a Speech Segment Network (SSN) approach is proposed for construction of a small speech unit set with which high quality speech can be synthesized. The SSN approach selects a speech unit set in which segmental and/or inter-segmental distortions are minimized by using combinatorial optimization methods such as iterative improvement or simulated annealing. Experimental results using diphone segments showed that the optimal diphone unit sets with total or maximum of inter-segmental distortion reduced by about 35%, 70% respectively can be constructed by this method. This reduction rate is enhanced as the segment population increased. Effectiveness of this unit set design was also perceptually confirmed by listening test using speech synthesized with the selected diphone unit set.

10.21437/ICSLP.1992-124

1 Introduction

In speech synthesis using concatenation of speech units, a well-designed unit set is of great importance for synthesized speech quality. The size of this set is expected to be small in a practical synthesis system, and a unit set which satisfies both these requirements is desired.

So far, when constructing such a unit set, desired unit set has been created by iterative replacement and listening (with heuristics). This operation is time-consuming, and has no guarantee that a better unit set can be obtained, because it is impractical to listen to all unit combinations. A method which selects the optimal speech unit set automatically is needed.

In concatenative speech synthesis, at least the following two different types of distortion have to be reduced to get high quality speech[1, 2].

a) **Segmental distortion:** for typicality of segment to corresponding context. Difference between spectral pattern of segment and the typical pattern for this target context.

b) **Inter-segmental distortion:** for smooth concatenation between segments. Difference of spectral pattern at concatenation point.

It is important to select a speech unit set which minimizes these distortions from a large speech database. Segmental distortion in a large database is measured by the sum of distances to all segments from the selected one in a cluster of similar contexts in the database. Inter-segmental distortion is measured by the sum of distances at a concatenation point between possible segments. For instance, in the case of a diphone unit set for synthesis, improving smoothness of concatenation particularly at vowel portions leads to higher quality synthesized speech.

As a method for constructing a unit set, the Contextual Oriented Clustering method[3] reduces segmental distortion. However, a method which reduces not only segmantal but also inter-segmental distortion is necessary for both high quality and small size of the unit set. The Speech Segment Network (SSN) approach which copes with this problem is described in the following section.

2 Proposed approach

2.1 Speech Segment Network

To minimize the above mentioned distortions, we considered the selection of a speech unit set as a combinatorial optimization problem. The cost value given to a unit set is minimized under the constraint that only one segment should be selected from each cluster to represent a phoneme sequence.

Speech Segment Networks (SSNs) are defined as networks of segments with values representing the degree of distortion that would occur both in each cluster and between the segments if concatenated in a synthesis process. A cost function, the sum of the segmental and/or inter-segmental distortions, is defined for each SSN. An optimal speech unit set can be obtained if all segments selected under the above constraints minimize this cost. The SSN is defined as a directed graph $G(V, E), E = E_1 \cup E_2$. Vertexes $v_i \in V$, edges $e_{i,j} = (v_i, v_j)$, (here, if $i = j$, then $e_{i,j} \in E_1$, else $e_{i,j} \in E_2$) are defined as following :

v_i : Each segments in a database

$e_{i,i} \in E_1$: Self loops which have values corresponding to segmental distortions $w_{i,i}$

$e_{i,j} \in E_2$: Edges which have values $w_{i,j}$ of inter-segmental distortions between vertexes which could be connected in a synthesis process.

For instance, an outline SSN for Japanese diphones consisting only of vowels is depicted in Fig.1 and Fig.2. For a sub network $C(V', E') \subset G$ which satisfies the above constraints in this network $G(V, E)$, a cost function

$$f(C) = \sum_{e_{i,j} \in E'} w_{i,j}$$

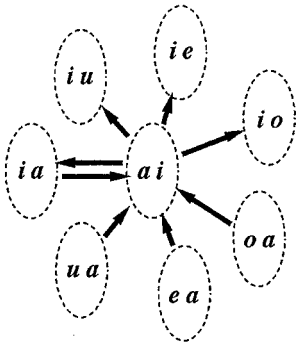


Fig.1 Links between cluster / a i / and others

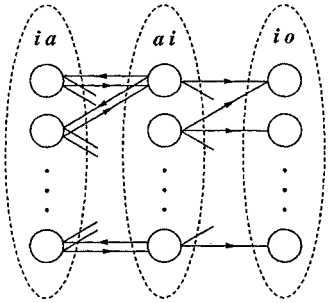


Fig.2 Links between vertexes for segments

is defined. The desired speech unit set which constructs C^* is obtained by searching C^* for the following:

$$f(C^*) = \min_C f(C)$$

This unit set has small segmental distortion and/or inter-segmental distortion. For searching C^*, the combinatorial optimization problem must be solved. A major advantage of the SSN approach is that it can reduce both segmental and inter-segmental distortions.

2.2 Optimization methods

Actually to solve the above search problem, all segments in a database are first put into clusters $V_i \subset V$ which are distinguished by phoneme contexts. The kind of clustering used depends on the kind of segments to be used as speech units for synthesis, i.e., syllable, diphone, triphone and non-uniform units. For instance, in the case of cv units, the number of cv clusters is about one hundred and twenty for Japanese. Generally though, more units need to be used for adequate speech quality.

Because the SSN optimization problem is NP-complete, as clusters or segments in each clusters increase, the number of combinations increases exponentially. It would take enormous computational time to solve this problem for checking all. As an efficient solution, methods which have been proved useful in other technical areas, such as branch-and-bound method, simulated annealing or iterative improvement, could all be used to solve this problem. We used an iterative improvement method and simulated annealing. These methods are easy to program, and can be considered to work well even if the size of the source database becomes very large.

3 Experiments for a diphone unit set

To evaluate the validity of the SSN approach, diphone unit sets for Japanese, consisting of two hundred and sixty-nine units, were selected by different methods. It was assumed that only one segment need be selected for each diphone string in the speech unit set. In the following experiment, only reduction of inter-segmental distortion by the SSN approach was evaluated because segmental distortion can be reduced easily, as mentioned later.

3.1 Segment network

ATR's isolated-word database[6] was used for a source of diphone units. There were many segments to choose from in each diphone cluster. The maximum number M_s of segments in a cluster was an experimental variable; 4, 6, 8 and 10 were tried, with the total number of segments being 990, 1441, 1879 and 2310 respectively. To create the SSN, values of the cepstral distance at each connection point between segments which could be connected in a synthesis process were given to edges in E_2, representing the inter-segmental distortion, or smoothness between segments. To check whether the SSN approach can reduce inter-segmental distortion in the speech unit set, zero values were given to edges in E_1.

3.2 Optimization methods

First of all, only segments which minimize distortion within clusters were selected ([**Centroid**] method) as a reference. This speech unit set minimizes sum of segmental distortion, but doesn't take care of inter-segmental distortion.

In experiments, both iterative improvement and simulated annealing were tested for optimization. Iterative improvement decreases the cost value deterministically and yields a local minimal value. Simulated annealing[4, 5] decreases the cost probabilistically and is able to achieve a global minimum. This technique is based on simulation of the annealing of solids. As temperature T decrease, the Boltzmann distribution concentrates on the states with lowest energy. The simulated annealing algorithm employed in experiments is following:

- Get an initial network C.
- Get initial temparature T_0.
- While not yet "frozen", do the following:
  - Repeat L times the following:
    - Pick a random cluster V_x
    - Select C_y which includes vertex v_y ($\in V_x$) with probability

$$\frac{e^{-f(C_y)/T}}{\sum_{v_i \in V_x} e^{-f(C_i)/T}}$$

    - Set $C = C_y$.
  - Update temparature T.
- Return C.

Actually, not only minimizing the sum of distortions, but also reducing their maximum is of importance. Therefore, unit sets were selected by the following methods in experiments.

[Min_Sum_II] By iterative improvement, networks were selected to minimize the sum of the inter-segmental distances.

[Min_Sum_SA] By simulated annealing, networks were selected to minimize the sum.

[Min_Max_II] By iterative improvement, networks were selected to minimize the maximal inter-segmental distortions using maximum norm.

$$\max_{i,j}\{w_{i,j}\} = \lim_{\eta\to\infty}\{\sum_{i,j} {w_{i,j}}^{\eta}\}^{1/\eta}$$

Actually, for simplicity of calculation, the minimization was carried out using the value $\sum_{i,j} {w_{i,j}}^4$. Network G' in which values of edges were ${w_{i,j}}^4$ was used instead of G.

[Min_Max_SA] By simulated annealing, networks were selected to minimize the maximum as above.

3.3 Results

SSN cost values which represent the sum of inter-segmental distortions, variances and maximum values in the selected unit sets are shown in Fig.3. **Min_Sum_II** and **Min_Sum_SA** methods reduced the sum of inter-segmental distortions by about $20 \sim 35\%$ of that achieved by the **Centroid** method(Fig.3-a). **Min_Max_II** and **Min_Max_SA** methods reduced the maximum value of inter-segmental distortions by about $50 \sim 70\%$ of that achieved by the **Centroid** method(Fig.3-b). These reduction rates got higher and absolute values smaller as M_s became larger. Variance by minimizing maximum was smaller than one by minimizing sum(Fig.3-c). It is hard to see a difference between **Min_Sum_II** and **Min_Sum_SA** but it was proved that **Min_Max_SA** has an advantage over **Min_Max_II** for reducing maximum inter-segmental distortion. The reason for this might be that it becomes hard to find a global minimum point of the cost function by larger variance of values of edges in SSN G' than G.

Distributions of inter-segmental distortion in unit sets by **Centroid**, **Min_Sum_SA** and **Min_Max_SA** in the case of $M_s = 10$ are shown in Fig.4. It was proved that distribution moved into lower area in inter-segmental distortion by **Min_Sum_SA** and **Min_Max_SA**.

For reference, distribution of total inter-segmental distortion of unit sets selected randomly 100,000 times are depicted in Fig.5. This figure shows that randomly selected unit sets have twice as much total inter-segmental distortion as those selected by **Min_Sum_SA**, and we can conclude that segments in these randomly selected unit sets will be difficult to connect smoothly.

Finally, the convergence process of the cost value by simulated annealing in **Min_Sum_SA**($M_s = 10$) is depicted in Fig.6 against values of temperature. It took about twenty hours to reach the minimum value on a DEC station 5000.

3.4 Perceptual evaluation test

A preference test was carried out, in order to evaluate whether speech synthesized with the unit set selected by the SSN approach was preferred over one selected by the **Centroid** method. Speech samples were synthesized by LMA filter[7] with thirty order cepstrum.

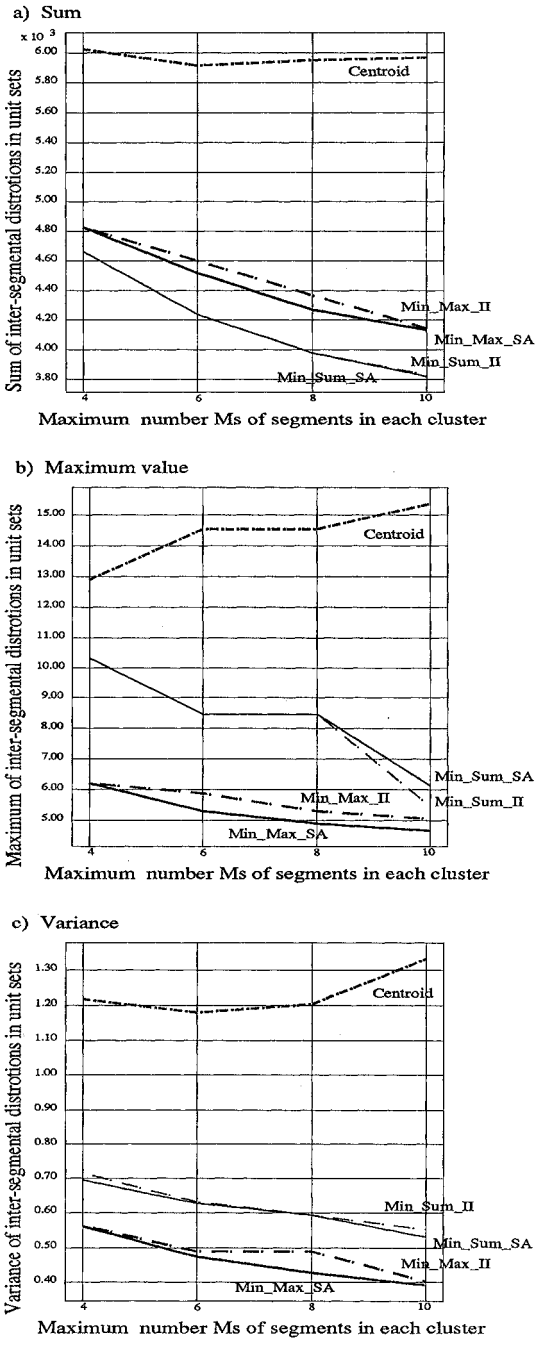


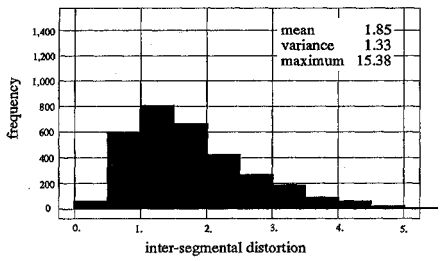
Fig.3 Statistic values of inter-segmental distortions in unit sets selected by different methods

[Subjects] 7 females

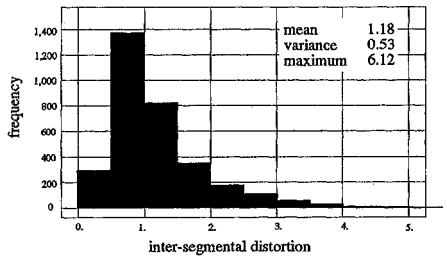
[Procedure] Judgment of "Which has the better quality?"

[Speech samples] A hundred words synthesized with unit sets selected by **A) Centroid** method, and **B) Min_Sum_SA** method in a case of $M_s = 10$.

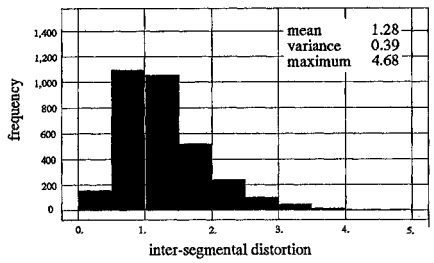
Many pairs of samples were hard to differentiate, and the overall preference scores didn't show clear differences. However,



a) Centroid method



b) Min_Sum_SA method



c) Min_Max_SA method

Fig.4 Distributions of inter-segmental distortions in unit sets selected by different methods (Ms = 10)

amongst those items which were clearly differentiated, i.e., selected by more than five out of seven subjects, **Min_Sum_SA** was preferred (**B** fourteen words, and **A** six). This shows that in a significant number of cases, **Min_Sum_SA** produces better quality output than the **Centroid** method.

4 Conclusion

An optimization method, Speech Segment Network (SSN) approach, for a speech unit set with which high quality speech can be synthesized, and yet which is small enough to be practical, has been proposed. Such a unit set is selected from a large database by minimizing both segmental and inter-segmental distortions simultaneously. This approach selects an optimal subset from the whole SSN made from a large database. Use of simulated annealing or iterative improvement methods overcomes the combinatorial difficulty.

Experimental results for selection of a diphone unit set showed that the SSN approach efficiently selects a unit set in which the sum and/or maximum of inter-segmental distortion is small. The synthesized speech quality from this unit set was good. A desired speech unit set can therefore be selected by the SSN approach automatically without any heuristic operations.

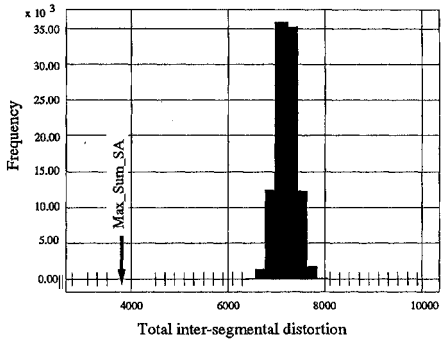


Fig.5 Distribution of total inter-segmental distortion at random trials (100,000 times)

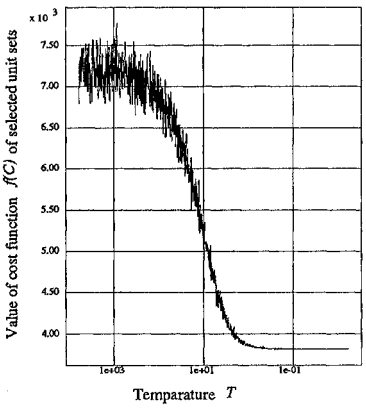


Fig.6 Decrease of cost at simulated annealing in Min_Max_SA (Ms = 10)

Acknowledgement.

Authurs are grateful to Masa-aki Sato in ATR Auditory and Visual Perception Research Labs. for helpful comments on simulated annealing.

References

[1] N.Iwahashi, N.Kaiki, Y.Sagisaka "Concatenative speech synthesis by minimum distortion criteria," *Proc. of ICASSP, Vol.II*, pp.65-68 (1992)

[2] K.Takeda, K.Abe, Y.Sagisaka "On the basic scheme and algorithms in non-uniform unit speech synthesis," *Talking Machines: Theories, Models, and Designs* Elsevier Science Publishers (1992)

[3] S.Nakajima, H.Hamada "Automatic Generation of Synthesis Units Based on Context Oriented Clustering," *Proc. of ICASSP*, pp.659-662 (1988)

[4] S.Kirkpatrick, C.D.Gelatt,Jr., M.P.Vecchi "Optimization by simulated annealing," *Science, Vol.220*, pp.671-680 (1983)

[5] P.J.M.van Laarhoven, E.H.L.Aarts "Simulated Annealing: Theory and Applications," D.Reidel Publishing Company (1987)

[6] K.Takeda, Y.Sagisaka, S.Katagiri "Acoustic-phonetic labels in a Japanese speech database," *Proc. of the European Conference on Speech Technology, Vol.2*, pp.195-198 (1987)

[7] S.Imai "Log Magnitude Approximation (LMA) Filter,"(in Japanese) *IEICE Vol.J63-A No.12*(1980)