ヒューマンインタフェース 44－2
（1992. 9. 10）

音声素片ネットワーク最適化による合成素片セットの構成法

岩橋 直人　　匂坂 芳典

ATR 自動翻訳電話研究所

本稿では、高品質な音声が合成される素片セットを、所望のデータ容量で構成する方法として、音声素片ネットワーク（ SSN : Speech Segment Network ）最適化法を提案する。本手法は、音声素片により構成されたネットワークに対してコスト関数を定義し、反復改善法やシミュレーテッド・アニーリング法などの組合せ最適化手法を用い、大容量データベースから、素片歪みと接続歪みを同時に最小化する音声合成素片セットを選択するものである。DIPHONE素片セットの構成実験を行った結果、本手法を用いることにより接続歪みの総和で約３５％、最大値で約７０％の低減が確認され、この効果は素片候補数が増すほど大きくなることが示された。また、受聴試験により音質向上に対する有効性が示された。

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

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% and 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 a listening test using speech synthesized with the selected diphone unit set.

1 Introduction

In speech synthesis using acoustical concatenation units, it is important to use suitable units for synthesized speech quality. Units are desired to represent allophonic characteristics properly and also be concatenated naturally. In the search for better speech quality, various types of units, such as syllable, diphone, triphone and non-uniform units [1], have been proposed.

In order to represent allophonic variation, the number of units shows a tendency to increase for different phoneme environments [2], and on reducing degradation at concatenation perceptually, concatenation at portions in consonants proved to be effective [3]. However, it is difficult to prepare all necessary phoneme combinations as a unit set.

Recently, using large unit database has proved to yield high quality speech to a certain extent, if a suitable unit selection procedure was employed [4, 5]. Even so in this type of synthesis system, the database should be constructed adequately so as to include suitable units phonemicaly and acoustically.

No matter what kind of units are used, a well-designed unit set is of great importance for synthesized speech quality. On the other hand, the size of this set is expected to be small in a practical synthesis system. It is desired that a unit set satisfies both these requirements on speech quality and its size.

Hitherto, when constructing such a unit set, the desired unit set has been created by iterative replacement and listening (with heuristics) with a large database. 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. An automatic procedure is needed for the selection of an optimal speech unit set.

To get high quality speech, the following two different types of distortion have to be reduced [5, 6].

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. As a method for constructing a unit set, the Contextual Oriented Clustering method[7] reduces segmental distortion. On the other hand, improving smoothness of concatenation leads to higher quality synthesized speech [4, 5, 6]. Therefore, 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. As the method which copes with this problem, the Speech Segment Network (SSN) approach is described in the following section.

2 Unit set construction by Speech Segment Network approach

2.1 Speech Segment Network

As a criteria for optimality of the unit set, segmental and inter-segmental distortions can be employed, because reducing these distortions in synthesized speech has a good effect on speech quality. Segmental distortion in a unit set 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.

To minimize these distortions in a unit set, 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.

To consider segmental and inter-segmental distortions in a unit set geometrically, 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-

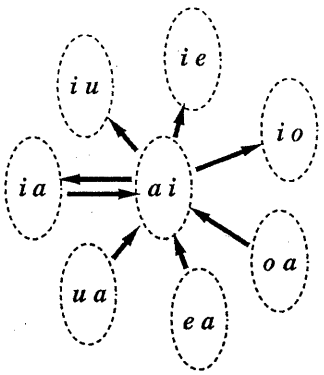

Fig.1 Links between cluster / a i / and others

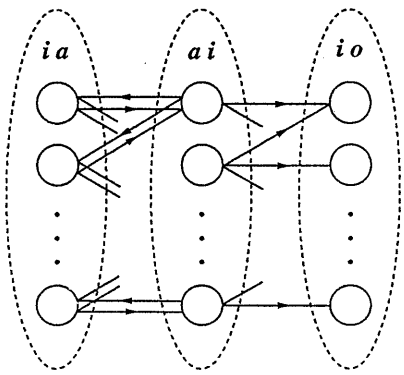

Fig.2 Links between vertexes for segments

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 shown 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

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

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 combinatorial optimization 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 cluster increase, the computational cost increases exponentially. Even if a cv unit set is used, it would take enormous time to solve the problem by checking all combinations. As an efficient solution, there are useful techniques, 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.

In iterative improvement, each selected combination is slightly changed at each step so that cost value decreases in iterative process. This

method decreases the cost value deterministically and yields a minimal value within a local search. Simulated annealing [8, 9] 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 can be implemented as following algorithm :

- 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.

Not only minimizing the sum of distortions, but also reducing their maximum is of importance. To minimize the maximal inter-segmental distortions, used is maximum norm;

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

and network G' in which values of edges are ${w_{i,j}}^{\eta}$ is used instead of G.

3 Experiments for a diphone unit set

Experiments were performed to evaluate whether a unit set could be selected by a SSN approach so as to reduce inter-segmental distortion in it. In the experiments, 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 Diphone Segment network

An isolated-word database[10] was used for a source of diphone units. This database consists of 5,456 words. 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. These values represent the inter-segmental distortion, or smoothness between segments. Inter-segmental distortions were caluculated at center portions of vowels, semivowels and nasals, and zero values were given at the others, because it is effective on speech quality to reduce inter-segmental distortions at former portions [6]. 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

First of all, a unit set which minimizes distortion within clusters was selected as a reference. This unit set can be obtained by choosing the nearest segment to a centroid in each cluster (**[Centroid]** method). 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. Unit sets were selected by the following methods;

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

[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. For simplicity of calculation, the minimization was carried out using the approximated 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.

3.3 Results

SSN cost values which represent the sum of inter-segmental distortions, maximum values and variances 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). Inter-segmental distortions became smaller as M_s became larger. Variance by minimizing maximum was smaller than one by minimizing sum (Fig.3-c). There isn't a large 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. This superiority of simulated annealing in minimizing maximum might be due to 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 are shown in Fig.4 for **Centroid**, **Min_Sum_SA** and **Min_Max_SA** in the case of $M_s = 10$. 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**.

Finally, the convergence process of the cost value by simulated annealing in **Min_Sum_SA** ($M_s = 10$) is shown in Fig.6 against values of temperature T. In this case, parameters in simulated annealing were set as $L = 6750, T_0 = 6000$, and T was updated by $T_{new} = 0.99 \cdot T_{old}$.

3.4 Perceptual evaluation test

A listening 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 [11] with thirty order cepstrum. Pitch pattern, phonemic duration and power component were kept same as natural utterances. And at concatened phonemes except for vowels and semivowels, automatic boundary adjustment [6] was carried out to minimize cepstral discontinuity.

[**Subjects**] 7 females

[**Procedure**] Preference judgment on ramdomly presented pairs of speech samples

[**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 almost same, and the overall preference scores didn't show clear differences. However, 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

Speech Segment Network (SSN) approach is proposed to obtain a speech unit set with which high quality speech can be synthesized, and yet which is small enough to be practical.

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 comparatively better than the unit set selected by usual method. A desired speech unit set can therefore be selected by the SSN approach automatically without any heuristic operations.

Because of independency on synthesis method, in any other synthesis scheme, e.g. PSOLA [12], the SSN approach is applicable.

Acknowledgement.

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

References

[1] Y.Sagisaka"Speech synthesis by rule using an optimal selection of non-uniform synthesis units," *Proc. of ICASSP IEEE* pp.679-682 (1988)

[2] 市川, 岩田, 三留, 伏木田 "規則合成における単位音声セットの検討," 電子情報通信学会 技術研究報告 SP87-6 (1987)

[3] 佐藤 "CVC と音源要素に基づく (SYMPLE) 音声合成," 音響学会 音声研究会資料 S83-69 (1984)

[4] T.Hirokawa,K.Hakoda"Segment selection and pitch modification for high quality speech synthesis using waveform segments", Proc. of ICSLP, pp.337-340 (1990).

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

[6] 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)

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

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

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

[10] 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)

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

[12] E.Moulines,F.Charpentier"Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication 9* pp.453-467 (1990)

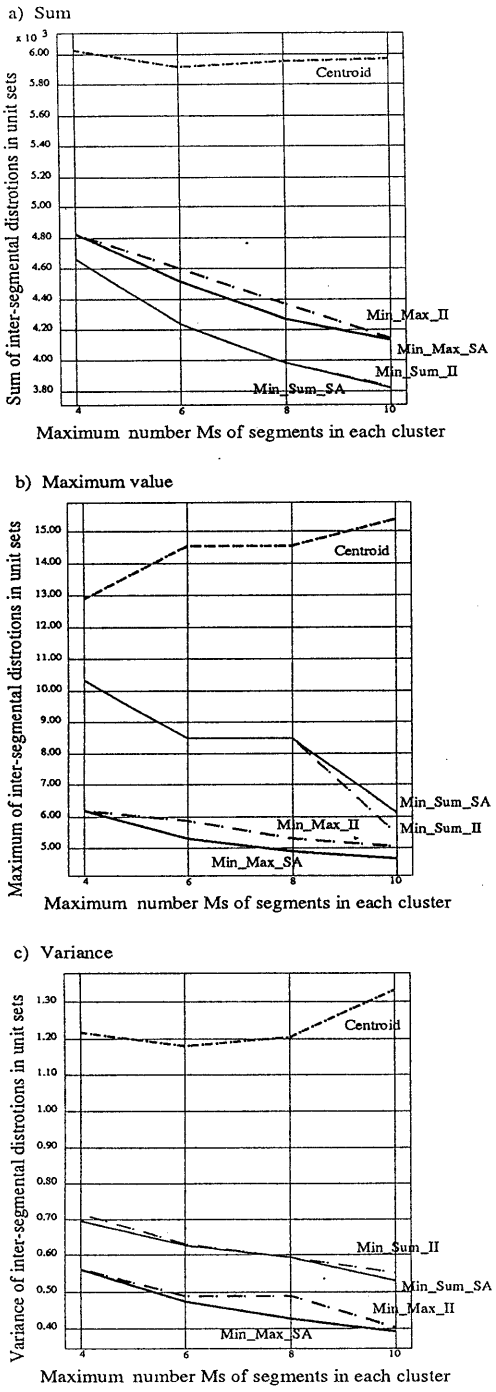

Fig.3 Statistic values of inter-segmental distortions in unit sets selected by different methods

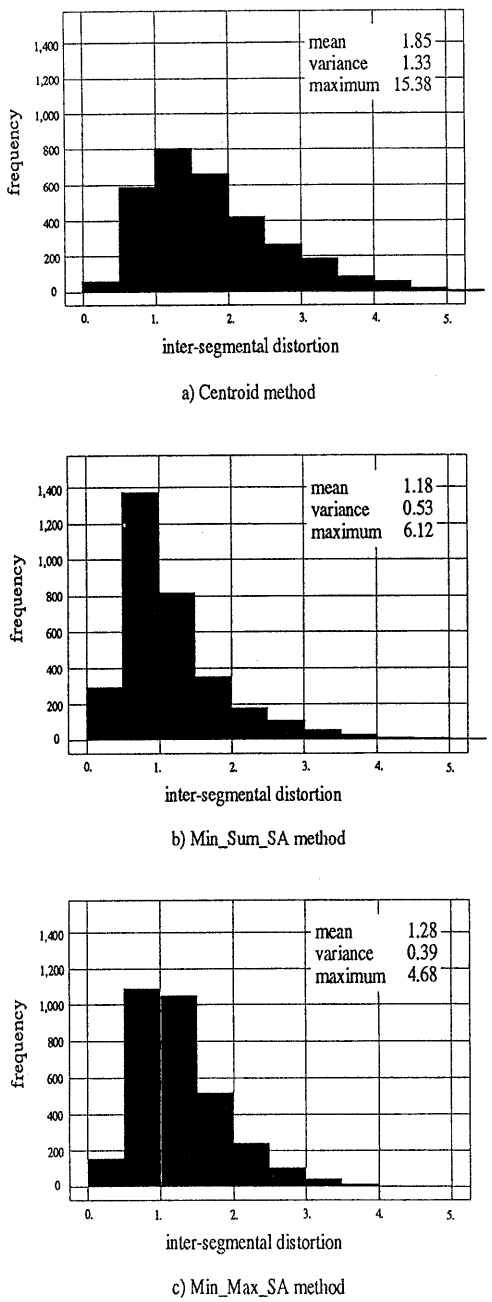

Fig.4 Distributions of inter-segmental distortions in unit sets selected by different methods (Ms = 10)

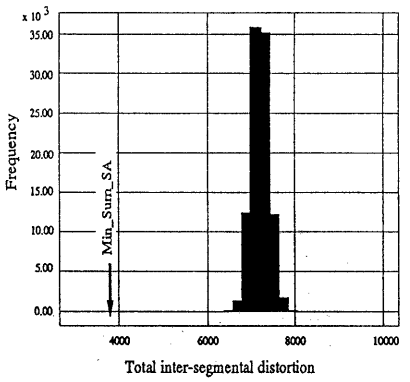

Fig.5 Distribution of total inter-segmental distortion at random trials (100,000 times)

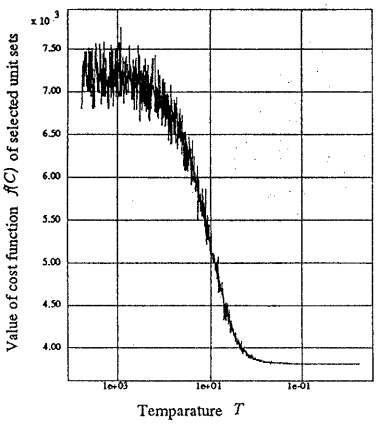

Fig.6 Decrease of cost at simulated annealing in Min_Max_SA (Ms = 10)