TR-IT-0021

規則音声合成の研究
A Study on Rule-based Speech Synthesis

岩橋 直人
Naoto Iwahashi

1993.9

概要

筆者が、1990 年 10 月 から 1993 年 9 月 までの 3 年間、自動翻訳電話研究所および音声翻訳通信研究所において行なった規則音声合成に関する研究について報告する。研究内容は、おおまかに以下の通りである。

- Segment selection procedure with mimimum distortion criteria
- Automatic labelling by HMM for synthsis segment
- Speech Segment Network approach for optimal synthsis unit set
- New multivariate analysis method and its application on duration modelling
- Speech individuality control with spectrum transformation by speaker interpolation

目 次

1 Abstract of research

筆者が、1990 年 10 月 から 1993 年 9 月 までの 3 年間、自動翻訳電話研究所および音声翻訳電話研究所において行なった規則音声合成に関する研究について報告する。研究内容の概略は以下のとおりである。

- Segment selection procedure with mimimum distortion criteria

  A new scheme is proposed for concatenative speech synthesis to improve the segment selection procedure by minimizing acoustic distortions between the selected segment and the desired spectrum for the target. The spectral prototypicality of a segment, the spectral difference between the source and target contexts, the degradation resulting from concatenation of phonemes, and the acoustic continuity between the concatenated segments are all considered as measures. A search method for selecting segments from a large speech database is also described. In this method, a three-step optimization is used for distortion minimization. A perceptual test shows that contextual spectral difference and acoustic continuity at the segment boundary are important measures for improving the quality of synthesized speech.

- Automatic labelling by HMM for generating synthsis segment

  単位連結型の規則音声合成に必要な音韻ラベリングの自動化についての検討を行なった。セグメンテーションシステムでは、最終的に適切な合成音声が出力されるような音韻ラベルが得られれば良く、この点からすると、NUU 音声合成方式を用いることで、セグメンテーションシステムへの要求を軽減できることが期待できる。合成システムでの使用可能性を見積もるために、自動化のための手法の一つとして考えられる連続出力分布型 HMM を用いた音韻セグメンテーションの実験を行なった。部分的に大きな誤りを含んではいるものの、全体的には規則合成システムでの使用が可能であると考えられるラベルが高い割合で得られた。また、得られたラベルに基づき、歪み最小化による NUU 音声合成方式を用いて合成音声を作成した。部分的に音質劣化が認められたが、セグメンテーション方式に音韻継続時間による拘束を付与することなどにより改善が可能であると考えられる。

- Speech Segment Network approach for optimal synthsis unt set

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.

- New multivariate analysis method and its application on duration modelling

  In this paper, statistical segmental duration modelling is proposed for English speech synthesis using Multiple Split Regression (MSR) and a hierarchical error function. To realize duration control by statistical method according to characteristics of English duration: interactions between control factors and hierarchical structure of timing, a suitable statistical modelling method is desired.
  MSR is a statistical modelling method which has data driven dynamic structure with combinatorial optimization technique. It incorporates both linear and tree regressions as special cases, and extends them. It can express phenomena of interaction between control factors for duration properly. The hierarchical predictive error function is adopted to analyze hierarchical structure of duration control in syllable and segmental levels.
  Experimental results show that MSR obtains higher values of multiple correlation than either linear or tree regressions with the same number of free parameters. Moreover, the error analysis by hierarchical predictive error function shows that interactions exist between factors at segmental and syllable levels in duration control, and that predictive errors at segmental duration are compensated in a syllable.

- Speech individuality control with spectrum transformation by speaker interpolation

A speech spectrum transformation method for interpolating spectral patterns between pre-stored speaker parameters for speech synthesis was proposed. The interpolation is carried out using Log Area Ratio parameters to generates the new spectrum. The formant structure can be transformed smoothly as the interpolation ratio is gradually changed, and speech individuality can be changed without degradation of speech quality.

Adaptation to a target speaker can be economically performed by this interpolation, which uses only a small amount of training data to produce a new speech spectrum sequence close to the target speaker's. An adaptation experiment was carried out using only one word spoken by the target speaker as data for learning showed that the distance between the target speaker's spectrum and the spectrum produced by our interpolation method is reduced by about 20% compared with distance between the target speaker's spectrum and spectrum of the speaker closest to the target speaker among pre-installed ones in the system.

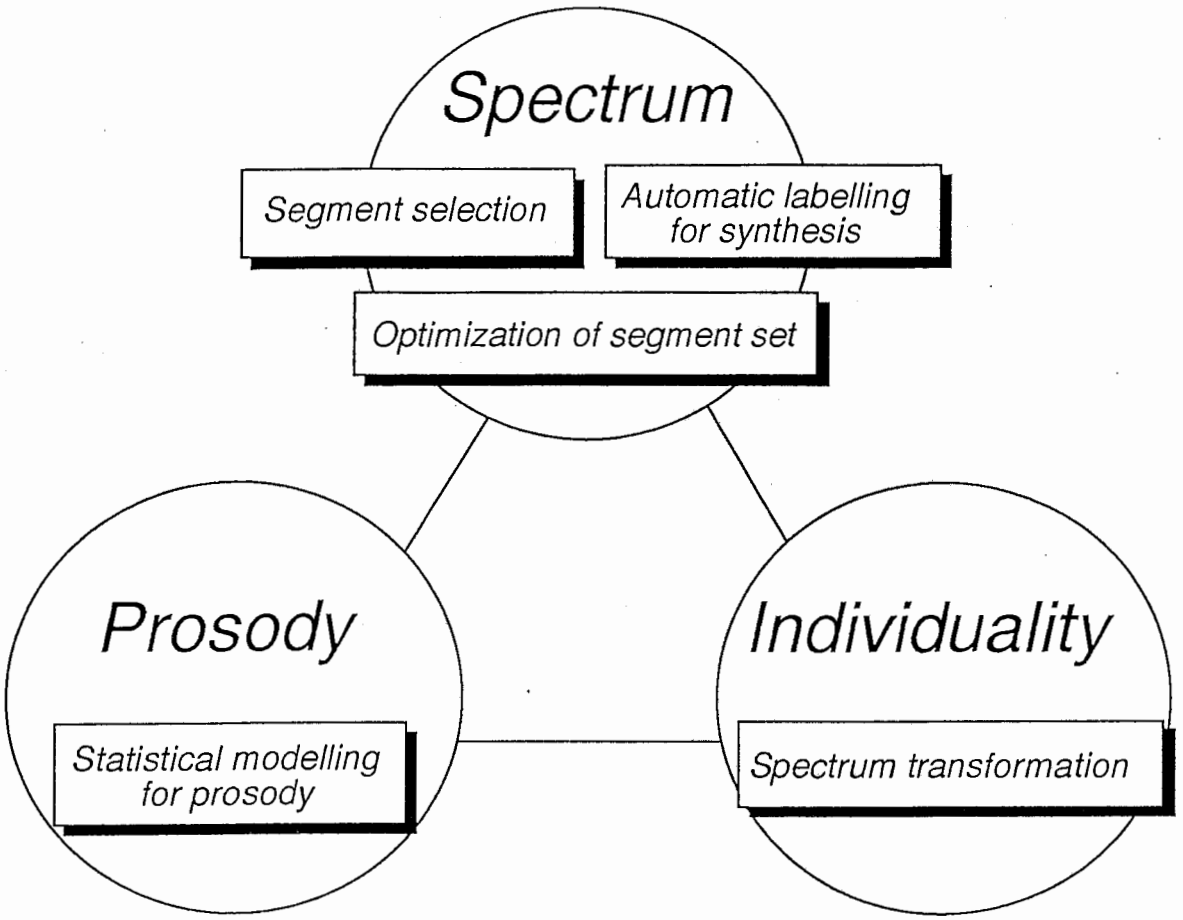

図 1: Researched Areas in Speech Synthesis

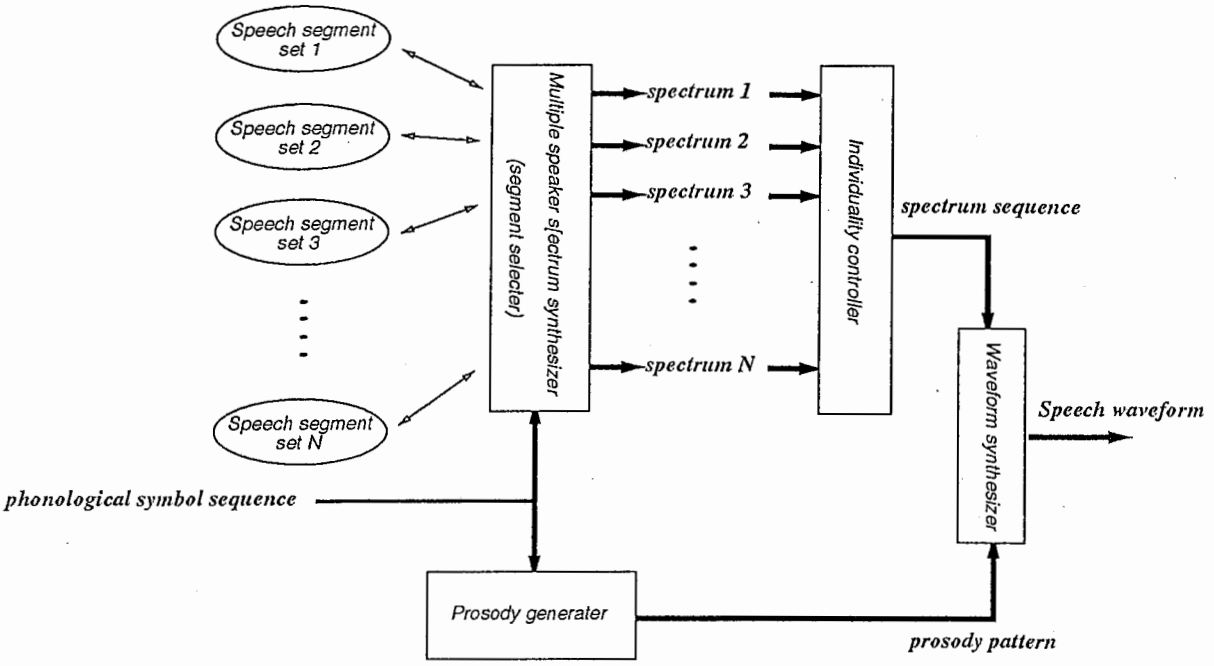

図 2: Targert synthesis system

2 Acknowledgement

Author is grateful to Dr. Sagisaka for his support and helpful discussion for three years. The accomplished research results cannot be obtained without him. Author is grateful to Dr. Campbell for impressive collaboration and support. Author is grateful to Dr. Kurematsu, Dr. Yamazaki and Mr. Sagayama for giving me chances of research activities. Author would like to thank researchers in ATR intrepreting telephony labs and interpreting telecommunications labs for their helpful discussions and comments.

3 List of papers

外部発表一覧

1. 岩橋 直人, 海木 延佳, 嵯峨山 茂樹, 匂坂 芳典: “音響的尺度を用いた合成単位素片の選択法,” 日本音響学会講演論文集, 1-6-1, pp. 207–208, (1991.3).
2. 岩橋 直人, 海木 延佳, 匂坂 芳典: “音響的尺度に基づく複合音声単位選択法,” 電子情報通信学会技術報告, SP91-5, pp. 33–40, (991.05).
3. 岩橋 直人, 藤原 紳吾, 小森 康弘, 杉山 雅英, 匂坂 芳典: “自動セグメンテーションによる音声合成単位の作成,” 日本音響学会講演論文集, 1-6-21, pp. 231–232, (1991.10).
4. 藤原 紳吾, 岩橋 直人, 小森 康弘, 杉山 雅英: “HMM とスペクトログラムリーディング知識に基づくハイブリッド音素セグメンテーションシステム,” 日本音響学会講演論文集, 2-5-20, pp. 87–88, (1991.10).
5. 岩橋 直人, 匂坂 芳典: “接続歪みを最小化する音声素片セットの構成法,” 日本音響学会講演論文集, 1-2-6, pp. 217–218, (1992.3).
6. 岩橋 直人, 匂坂 芳典: “歪み最小化音声合成手法の主観・客観評価,” 日本音響学会講演論文集, 2-2-15, pp. 281–282, (1992.3).
7. Naoto Iwahashi, Nobuyoshi Kaiki, Yoshinori Sagisaka: “Concatenative Speech Synthesis by Minimum Distortion Criteria”, *International Conference on Acoustics, Speech, and Signal Processing*, (1992.3).
8. 岩橋 直人, 匂坂 芳典: “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,” 情報処理学会ヒューマンインターフェース研究会, (1992.9).
9. 岩橋 直人, 匂坂 芳典: “空間分割型数量化 I 類による音声制御の統計モデリング,” 日本音響学会講演論文集, (1992.10).
10. 岩橋 直人, 匂坂 芳典: “音声素片ネットワーク最適化法法による DIPHONE 素片セットの構成,” 日本音響学会講演論文集, (1992.10).
11. 王文俊, Nick Campbell, 岩橋 直人, 匂坂 芳典: “Unit Selection for English Speech Synthesis Using Regression trees,” 日本音響学会講演論文集, (1992.10).
12. Naoto Iwahashi, Yoshinori Sagiska:“Speech Segment Network approach for an optimal synthesis unit set,” *International Conference on Spoken Language Processing*, (1992.10)
13. 岩橋 直人: “論文紹介： The estimation of stochastic context free grammars using the Inside-Outside algorithm”, 情報処理学会誌 (1993.1).
14. 岩橋 直人, 匂坂 芳典: “空間多重分割型数量化法による英語音声のセグメント継続時間制御モデル”, 日本音響学会講演論文集, (1993.3).
15. 三村 克彦, 岩橋 直人, 匂坂 芳典: “単位接続歪みが合成音明瞭度に与える影響について”, 日本音響学会講演論文集, (1993.3).
16. Naoto Iwahashi, Yoshinori Sagisaka: “Duration Modelling with Multiple Split Regression”, *Proceedings of EUROSPEECH '93*, (1993.9)
17. 岩橋 直人, 匂坂 芳典: “話者内挿処理による声質制御法”, 日本音響学会講演論文集, (1993.10).

18. 平井 俊男, 岩橋 直人, Helen Valbret, 樋口 宣男, 匂坂 芳典: "統計的手法による基本周波数パタンの制御", 日本音響学会講演論文集, (1993.10).

19. Naoto Iwahashi, Yoshinori Sagisaka: "Speech segment selection for concatenative synthesis based on spectral distortion minimization", Trans. of IEICE (1993.12)

20. Naoto Iwahashi, Yoshinori Sagisaka: "Speech Segment Network approach for synthesis unit set", (投稿中)

21. Naoto Iwahashi, Yoshinori Sagisaka: "Speech spectral transformation by speaker interpolation", *International Conference on Acoustics, Speech, and Signal Processing* (投稿中).

22. Naoto Iwahashi, Yoshinori Sagisaka: "Multiple Split Regression and its application for duration modelling", (投稿予定)