氏　　名	TRUNG NGHIA PHUNG				
学位の種類	博士(情報科学)				
学位記番号	博情第 285 号				
学位授与年月日	平成 25 年 9 月 24 日				
論文題目	Studies on Improving the Naturalness of Synthesized Speech under Limited Data Conditions (制限されたデータのもとでの合成音声の自然性向上法に関する研究)				
論文審査委員	主査	赤木　正人	北陸先端科学技術大学院大学		教授
		党　　建武		同	教授
		鵜木　祐史		同	准教授
		田中　宏和		同	准教授
		北村　達也	甲南大学		教授

論文の内容の要旨

The motivation of this dissertation was to propose methods for improving the naturalness of synthesized speech under limited data conditions.

Because speech is the result of sequential linking of phonetic units, a speech synthesizer requires a database that covers all phonetic units in a specific unit set to synthesize any input text content, resulting in a requirement of significant amount of data for synthesizing (B. Bozkurt and T. Dutoit, 2003).Due to the efforts of co-articulation on speech synthesis (SS), not only all context-independent phonetic units but also all context-dependent phonetic units, are necessary to synthesize natural speech. As a result, state-of-the-art speech synthesizers require large-scaled speech corpora to synthesize natural speech (J. Kominek and A. Black, 2003; D. Suendermann and A. Black, 2010).

Building a large-scaled speech corpus is a costly task that takes a long time and a great deal effort by engineers, acousticians and linguists. Therefore, high-quality SS under limited data conditions is important in practice, specifically for under-resourced languages. Synthesizing speech with a limited amount of data is also critical for customizing synthesized speech with multiple voices (T. Toda .et.al, 2007). It is also critical in movies and games applications when we need to synthesize the voices of whom are not alive or have lost their voice characteristics (O. Turk, 2007). Since methods of voice transformation can be used with a limited amount of target data, a target voice can be synthesized by using a two-step procedure, which synthesizes a standard voice with a large-scaled database first and transforms the synthesized standard voice to the target voices later (T. Toda .et.al, 2007). However, this approach still requires a large-scaled database for the first step of synthesizing the standard voice. ***As a result, to directly build highly- natural speech***

synthesizers under limited data conditions is an important and interesting research topic.

In the literature, there are a few approaches to directly improve the naturalness of synthesized speech under limited data conditions. The first approach is to maximize the use of existence contexts in the database (Y. Sagisaka, 1988; K. Tokuda and A. Black, 2002). This approach has been shown its efficiency compared with traditional approaches. However, SS with this improvement still requires large amounts of original data for concatenation or training. The second approach is to use methods of speech modification to reduce mismatch-context errors occurred when phonetic units in matching contexts are not available due to the limitation of the database (D. Chappell and J. Hansen, 2002; A. Kain and J. Van Santen, 2007). This approach has just used in CSS and has shown a little improvement.

One core problem in synthesizing natural speech under limited data conditions is to en- sure the smoothness of synthesized speech. Both temporal and spectral over-smoothness and over-roughness can reduce the naturalness of synthesized speech. Under limited data conditions, the over-roughness in speech synthesized by concatenative SS (CSS) (A. Kain and J. Van Santen, 2007) and the over-smoothness in speech synthesized by Hidden- Markov-Model-based SS (HMMSS) (T. Toda and K. Tokuda, 2007) increase. This problem significantly reduces the naturalness of synthesized speech under limited data conditions. However, ensuring an "appropriate smoothness" for synthesized speech has not taken deeply into consideration in research fields of SS. ***Therefore, the unified purpose of this research is to use methods of speech modification and transformation to ensure an "appropriate smoothness" in synthesized speech under limited data conditions.***

Temporal decomposition (TD) (B.S. Atal, 1983) is an interpolation method decomposing a spectral or prosodic sequence into its sparse event targets and correspondent temporal event functions. The modified restricted TD (MRTD) (C. Nguyen and M. Akagi, 2003) is one simplified but efficient version of TD. With a determination of event functions close to the concept of co-articulation in speech, MRTD can synthesize smooth speech. The smoothness in synthesized speech can be adjusted by modifying event targets of MRTD (B. Nguyen and M. Akagi, 2009). Therefore, MRTD can be used to modify / transform / and synthesize speech with an "appropriate smoothness". As a result, MRTD was used throughout the proposed methods in this research.

Constructing SS under limited data conditions is more practical for under-resourced languages, where large public speech corpora are missing. As the tonal-monosyllabic Vietnamese is an under-resourced language (C.M. Luong, 2006), Vietnamese datasets were used in this paper for implementations and evaluations.

Based on the general motivation and the unified purpose of this research, four specific objectives were specified and solved.

The first objective was to propose a new and efficient speech smoothness measure to control and to evaluate the smoothness in synthesized speech. A speech smoothness measure based on the square sum of the variance of the delta-delta sequence, named distance of global smoothness measure (DGSM), in both time and spectral domains was proposed. The proposed DGSM was shown to be reliable and efficient to measure the smoothness in different kinds of speech in both time and spectral domains.

The second objective was to solve the main problem of CSS under limited data conditions, which is reducing mismatch-context errors in CSS that cause the temporal over- roughness in synthesized speech. A model of contextual effect in CSS, a method of speech modification, and a method of unit selection were proposed and were combined into a non-uniform CSS to reduce mismatch-context errors in CSS under limited data conditions. MRTD was used as a core of these proposed methods to obtain synthesized speech with an "appropriate smoothness". Experimental results with Vietnamese datasets show that mismatch-context errors could be solved and the proposed CSS was natural in terms of smoothness under limited data conditions.

The third objective was to solve the main problem of HMMSS under limited data conditions, which is reducing both the temporal and spectral over-smoothness in HMMSS. A hybrid SS between HMMSS and MRTD, referred to as HTD, was proposed to reduce over- smoothness in synthesized speech under limited data conditions. The two components event functions and event targets of MRTD were independently controlled to obtain an "appropriate smoothness" in synthesized speech. Experimental results with Vietnamese datasets show that speech synthesized by the proposed HTD had a more "appropriate smoothness", compared with that by the state-of-the-art HMMSS using GV, resulting in an improvement on the naturalness in synthesized speech.

The fourth objective was to solve one case with ultra-limited data conditions for tonal languages when the number of tonal units is not sufficient by using a method of tone transformation. Lexical tones are usually represented by fundamental frequency (F0) contours. Therefore, a tone transformation can be considered as a F0 contour transformation applied for converting lexical tones. A method of F0 transformation using MRTD and GMM was proposed to ensure the "appropriate smoothness" in the transformed F0 contours of the lexical tones. Experimental results with Vietnamese datasets show the effectiveness of the MRTD-GMM F0 transformation and it could be applied to improve the usability of SS of tonal languages under limited data conditions.

In summary, methods of speech modification and transformation were proposed to improve the naturalness of synthesized speech under limited data conditions based on the concept "appropriate smoothness" in synthesized speech.These methods showed their efficiencies on improving the usability under limited data conditions for both CSS and HMMSS. There results contribute to the research fields of speech

processing by introducing the new concept "appropriate smoothness" in speech. There results also contribute to the research and development fields of SS in order to make SS more convenient and more efficient for human-machine interaction systems.

論文審査の結果の要旨

本論文は、少量の音声資源からの自然性の高い合成音声作成法に関する研究の報告である。現在、広く用いられている音声合成法、たとえばユニット結合型音声合成法、あるいは、HMM 音声合成法では、大量の音声資源を用意できる場合には、合成音声の自然性および了解性は、ほぼヒトが発した音声に匹敵するレベルまで向上している。しかし、少量の音声資源しか用意できない場合（たとえば、少数民族の言語、発展途上国の言語、すでに死に絶えた言語等）は、自然性の中の大きな因子である「滑らかさ」（smoothness）が適切に制御されず、自然性および了解性が低下する。特に、ユニット結合型音声合成法では、ユニットの個数が制限されるため、適切なユニットペアが選択されず「滑らかさ」が粗くなる（over-roughness）。また、HMM 音声合成法では、尤度の高い音声を少ない資源から推定するため滑らかすぎる（over-smoothness）音声が合成される。

本論文では、自然性の中の大きな因子である「滑らかさ」（smoothness）を適切に制御するという問題に対して、まず、最適な滑らかさ（appropriate smoothness）を定式化している。そして、Temporal Decomposition (TD)による音声特徴の分解と、分解された音声特徴への Voice Conversion (VC)による変形手法を適用して、ユニット結合型音声合成法に対しては結合部の滑らかさを担保する、また，HMM 音声合成法に対しては音声特徴の変化を強調する。これにより smoothness を最適な値である appropriate smoothness に近づけることを可能としている。具体的には、

(1) appropriate smoothness 測度の定義：現在、「滑らかさ」の測度として多く用いられているのは、音声特徴の変化に対する global variance (GV)である。しかし、GV では変化の速度を適切に表現できない。本論文では、「滑らかさ」の測度として音声特徴の変化に対する曲率の二乗和を定義した。これにより、合成音声の over-roughness および over-smoothness が計測でき滑らかさを最適化するための土台ができた。

(2) ユニット結合型音声合成法での over-roughness の回復：ユニット結合型音声合成法では、音声資源を音素、音節などの小さいユニットに分割しておき、合成する文章に合わせて、適切なユニットペアを選択・結合する。音声資源が少量である場合には、適切なユニットペアを選択できず、結合部で音響特徴の変化が不連続となり、over-roughness の印象を与える。本論文では，TD により各音声ユニットを音声イベントとその補間関数に分解する。そして、結合時に最適な「滑らかさ」を得られるように音声イベントを VC により変形した上で結合する。音声資源が少量であるベトナム語をターゲットとして音声を合成した結果，提案法は現有のユニット結合型音声合成法に比べて、より自然な合成音を生成した。

(3) HMM 音声合成法での over-smoothness の回復：HMM 音声合成法では，大量の音声資源から音響モデルを統計的にあらかじめ学習し、尤度の最も高い音響特徴の軌跡を合成に使用する。音声資源が少量である場合には適切に学習が進まないため、生成された音響特徴の軌跡はしばしば滑らかすぎる（over-smoothness）こととなる。本論文では、HMM 音声合成法と TD のハイブリッド合成器を提案する。HMM 音声合成法で一旦合成された音声に対して TD を適用することにより、音声イベントとその補間関数に分解する。over-smoothness の原因は、主に音声イベントの歪であるため、VC により音声イベントを修正変形した上で TD により再合成する。ベトナム語をターゲットとして音声を合成した結

果、最適な値である appropriate smoothness に近づけることができた。

これらの手法により、最適な滑らかさに制御された合成音声は、少数の音声資源からの合成音声であるにもかかわらず、自然性が現有のものより向上し、了解性は保たれることがわかった。

以上のように、本研究は新しい概念のもとで、少数の音声資源からの合成音声の自然性改善を行い、聴取実験の結果従来法よりもより自然な音声合成を実現したものであり、学術的に貢献するところが大きい。よって博士（情報科学）の学位論文として十分価値あるものと認めた。