음성신호에 대한 WSOLA 알고리즘 기반 시간축변환의 계산량 감축

Complexity Reduction of Time-Scale Modification of Speech Based on Waveform Similarity Overlap-and-Add

김덕수, Duk Su Kim*, 이영한, Young Han Lee**, 김홍국, Hong Kook Kim***, 김명보, Myeongbo Kim**, 김상룡, Sang Ryong Kim**

요약 본 논문에서는 waveform similarity overlap-and-add (WSOLA) 알고리즘 기반 시간축변환(time-scale modification; TSM)의 처리시간 감축을 위해, 음성신호의 주기(period)를 기반으로 최적이동지점(optimal shift)을 예측하여 상호상관도(cross-correlation) 연산을 감소하는 방법을 제안한다. WSOLA 알고리즘은 출력신호를 결정하기 위해 매 프레임마다 기준신호(reference signal)와 검색범위(search range) 내의 후보신호(candidate signal)간 상호상관도 연산을 수행한다. 하지만 음성신호는 단구간 내에서는 피치 주기로 반복되기 때문에 기준신호와 후보신호 중 선택된 신호(selected signal) 간의 시간차는 피치 주기와 유사한 값을 갖는다. 따라서, 과거 프레임에 대해 상호상관도 연산을 수행해 얻은 추정 주기를 이용하여 현재 프레임의 상호상관도 검색범위를 감축할 수 있다. 구체적으로, 홀수 프레임에 대해서는 전체범위에 대해 상호상관도 연산을 수행하고, 짝수 프레임에 대해서는 추정 주기를 이용해 최적이동지점을 예측하여 상호상관도 검색범위를 줄인다. 실험 결과, 본 논문에서 제안하는 방법은 모든 프레임에서 전체범위에 대해 상호상관도 계산을 하는 방법에 비해 비슷한 수준의 음질을 유지하면서 TSM 처리 계산량을 15% 정도 감축하였다.

Abstract In this paper, we propose a complexity reduction technique for time-scale modification of speech signals based on a waveform similarity overlap-and-add (WSOLA) algorithm. The proposed technique tries to easily estimate an optimal shift of WSOLA for the current frame using the shift obtained from the previous frame. This is based on the fact that a short-time speech signal is periodic with a pitch period, thus the time difference between the reference signal and a selected signal for WSOLA would be similar to the pitch period. Therefore, we can reduce the complexity required for computing cross-correlations which are the main computational burden for WSOLA. It is shown from time-scale modification experiments that the proposed technique has a comparable speech quality with 15 % complexity reduction, compared to the full search for WSOLA.

핵심어: Time-Scale Modification, Waveform Similarity Overlap-and-Add (WSOLA), Complexity Reduction, Period Estimation

*주저자 : 광주과학기술원 정보통신공학과 석사과정 e-mail: dskim867@gist.ac.kr

**공동저자 : 광주과학기술원 정보통신공학과 박사과정 e-mail: cpumaker@gist.ac.kr

삼성전자주식회사 캠코더사업팀 수석연구원 email: kmbo.kim@samsung.com

삼성전자주식회사 캠코더사업팀 전무이사 email: srkim@samsung.com

***교신저자 : 광주과학기술원 정보통신공학과 교수 e-mail: hongkook@gist.ac.kr

1. 서론

시간축변환(time-scale modification; TSM)은 음성의 길이를 조절하는 기술로, 다양한 분야에서 활용이 되고 있다. 예를 들어, TSM 을 음성인식의 전처리과정에 사용함으로써 인식률을 향상시킬 수 있으며[1], 음성합성에서는 음의 길이(duration)를 조절하는데 쓰이고 있다[2]. 이러한 시간축변환 기술은 자원이 부족한 소형 정보기기 환경에서 구현되기 위해서는 계산량 감축이 필수적이다. 기존에는 계산량을 감축시키기 위해 상호상관도 계산 구간을 표본화하여 전체 계산량을 감축시키는 방식이 제안되었다[3]. 하지만 기존의 방식 모두 최대 상호상관도를 갖는 위치를 정확하게 찾지 못한다는 단점을 가지고 있기 때문에 음질 열화를 야기할 수 있다.

본 논문에서는 waveform similarity overlap-and-add (WSOLA)를 기반으로 하는 TSM 수행하는 데 있어서, 음질의 열화 없이 WSOLA 의 계산량을 감축하기 위해 피치 기반의 최적이동지점 예측을 통하여 상호상관도 계산의 검색범위를 줄이는 방법을 제안한다.

2. WSOLA 알고리즘 기반 TSM

WSOLA 알고리즘 기반 TSM 은 overlap-and-add (OLA) 알고리즘 기반 TSM 이 가지고 있는 위상왜곡 현상을 최소화하여 음질을 향상시키기 위해 개발되었다[4]. <그림 1>은 2 배속일 경우에 대한 WSOLA 알고리즘의 동작과정을 보여 준다. 그림에서 L 은 OLA 구간의 길이를 의미하며, 신호 (1)은 k 번째 프레임의 출력을 의미하며, (k+1)번째 프레임의 출력을 생성하기 위해서 기준신호(reference signal) (1')과 (2)를 중심으로 $-\Delta_{\max} \sim +\Delta_{\max}$ 범위내의 후보신호(candidate signal) 간의 상호유사도를 계산한다. 마지막으로 이 중 최대값을 갖는 출력신호 (2')를 결정한 후 출력신호 (1)와 OLA 방식으로 합성한다.

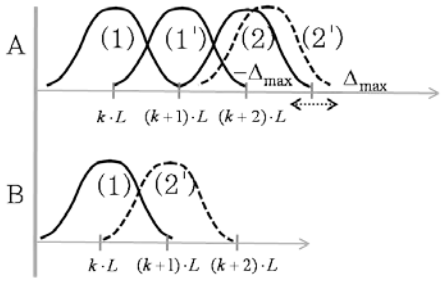

그림 1. WSOLA 알고리즘 동작과정

3. 제안된 피치 예측기반 계산량 감축 방법

음성신호는 피치 주기에 따라 신호가 반복되는 특성을 가진다. 따라서 음성 신호에 대한 WSOLA 알고리즘 동작 시, 다음 식 (1)과 같이 기준 신호와 최적이동지점의 거리로 음성의 주기를 예측할 수 있다.

$$P(k+1) = L \cdot (1-\alpha) + \Delta(k-1) - \Delta(k) \qquad (1)$$

여기서, $P(k+1)$ 는 (k-1) 번째 및 k 번째 프레임으로 부터의 예측된 (k+1)번째 주기, α 는 배속, $\Delta(k)$ 는 k 번째 프레임의 최적이동지점 의미한다. 여기서 구한 주기는 다음 프레임의 최적이동지점을 예측하기 위해 사용한다. 즉,

$$\Delta(k+1) = \arg\min_{l} \left| (k+1) \cdot L \cdot \alpha - \beta \right| \qquad (2)$$

여기서, $\beta = (k \cdot L \cdot \alpha + \Delta(k) + L) - l \cdot P_E$ 이고 $\Delta(k+1)$ 는 (k+1)번째 프레임의 예측된 최적이동지점을 나타낸다. 따라서, 예측된 최적이동지점을 기준으로 작은 범위 $-\Delta_s \sim +\Delta_s$ 에 대해 상호상관도 계산을 함으로써 계산량을 감축할 수 있다. 홀수 프레임에 대해서는 검색범위를 전체로 하고, 짝수 프레임에 대해서는 홀수 프레임 알고리즘 수행 시 예측한 최적이동지점을 이용해서 상호상관도 계산의 검색범위를 줄일 수 있다.

4. 실험 및 성능 평가

실험을 위해 16 kHz 모노 음원을 사용하였다. 본 논문에서 제안하는 방법을 적용하면 기존 시간축변환의 계산량에 비해서 15% 정도 감축되는 효과가 있었다. 또한 제안하는 방법을 적용하였을 때 음질이 떨어지지 않는 것을 증명하기 위해 음질평가를 수행하였다. 음질평가에는 청각에 이상이 없는 남녀 8 명이 참여하였다. <표 1>은 음질평가 결과를 보여 준다. 표에서 보는 바와 같이, 계산량을 감축하여도 음질이 유사한 것을 확인할 수 있었다.

방식	기존 방식	유사	제안한 방식
선호도	7.5%	92.5%	0.0%

표 1. 2 배속으로 처리된 음원에 대한 AB 선호도 테스트 결과

5. 결론

본 논문에서는 추정 주기기반의 최적이동지점 예측을 이용한 WSOLA 알고리즘의 계산량 감축 방법에 대해 제안하였다. 음성신호는 주기마다 반복하는 특징이 있기 때문에, WSOLA 알고리즘 적용 시 음성의 주기를 예측할 수 있고 이를 이용해서 다음 프레임에 대한 최적이동지점을 예측할 수 있었다. 본 논문에서 제안하는 방법을 적용하였을 때 음성의 주요 특징인 피치를 보존하면서 계산량을 15% 정도 감축시키는 효과가 있었다.

참고문헌

[1] N. R. Chong-White and R. V. Cox, "Enhancing speech intelligibility using variable-rate time-scale modification," *U.S. Patent 7,065,485*, 2006.

[2] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, nos. 5-6, pp. 453-467, Dec. 1990.

[3] J.-H. Chen, "Audio time scale modification using decimation-based synchronized overlap-add algorithm," *U.S. Patent Application 20070094031*, 2007.

[4] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. ICASSP*, Minneapolis, MN, pp. 554-557, Apr. 1993.