MFCC 특징 추출을 위한 Floating-Point 비트 너비 최적화 방안

*조지혁, 유호영, 차소영, 박인철
한국과학기술원 전기및전자공학과
e-mail : *jhjo.ics@gmail.com, hyyoo.ics@gmail.com, sycha.ics@gmail.com, icpark@kaist.edu*

Optimization of Floating-Point Bit-width for MFCC Feature Extraction

*Jihyuck Jo, Hoyoung Yoo, Soyoung Cha, In-Cheol Park
Department of Electrical Engineering
KAIST

Abstract

Earlier research efforts have been focused on fixed-point mel-frequency cepstrum coefficients (MFCCs) [1][2], but such a number representation could not fulfil the full dynamic range of the feature extraction process with small bit-width. In order to overcome the limitation, an optimization algorithm of floating-point bit-width for MFCC feature extraction process is examined in this paper. The proposed algorithm is used in speech recognition system simulator. As a result, the optimized floating-point speech recognition system simulator achieves 4.16 times reduction in bit-width without significant degradation of the word correction rate.

I. 서론

음성인식 시스템은 대표적인 인간 컴퓨터 인터페이스로, 오늘날 스마트폰, TV, 자동차 등의 다양한 분야에서 활용되고 있다. 일반적으로 상기 음성인식 시스템은 web-based 음성인식 혹은 digital signal processor를 기반으로 한 음성인식을 사용한다[3]. 대용량의 데이터가 처리되는 음성인식 과정의 특성 상, 위와 같은 음성인식 방식은 최소의 전력 소모가 요구되는 모바일, 임베디드 시스템 등에 적합하지 않다.

이를 보완하기 위하여 최근 연구에서는 음성인식 과정을 저전력 하드웨어로 구현하고자 노력을 해왔다. 음성인식 시스템은 크게 음성 신호로부터 특징 벡터를 추출하는 특징 추출 (feature extraction) 과정과 상기 벡터를 해석하고 인식하는 음성 분류 (classification) 과정으로 나뉜다. 특징 추출 과정에는 linear predictive coding 혹은 mel-frequency cepstrum coefficients (MFCCs) 등이 있는데, 이 중 계산량 대비 정확도가 높은 MFCC가 주로 사용된다. [1]에서는 음성인식 과정 중 MFCC 추출 과정에 대하여 전력 소모량을 감소하기 위해 공유하는 fixed-point 가감기 및 곱셈기를 사용하였다. [2]는 음성인식 시스템의 모든 fixed-point 연산에 대하여 비트 너비를 최적화하였다. 하지만 위의 논문들과 같이 fixed-point 연산으로 dynamic range가 큰 음성인식 시스템을 구현하는 것은 전력 소모 측면에서 부적합하다.

이에 대한 해결책으로, 본 논문에서는 음성인식 과정 중 데이터 처리가 많은 MFCC 과정에 대하여 비트 너비가 최적화된 floating-point 연산을 적용하는 방안을 제시한다. 이를 통하여 fixed-point 대비 대부분의 연산자의 비트 너비가 2배, 최대 4.16배 감소되었다. 결과적으로 해당 연산에 대하여 combinational logic의 비트 너비가 줄어들고 연산에 사용되는 전력 소모를 최소화할 수 있다.

본 논문의 구성으로, 본문 II에서 floating-point 연산의 비트 너비를 최적화하는 알고리즘에 대하여 제안하고, 본문 III에서 상기 연산을 적용한 음성인식기 동작을 보이며, 본문 IV에서 결론을 맺는다.

II. 본론

2.1 MFCC 특징 추출 과정 및 특성

MFCC 특징 추출 과정은 음성신호처리에서 많이 사용되는 특징 추출 방법 중 하나이다. Mel-frequency cepstrum (MFC)이란 20ms 가량의 단위 시간 (frame) 내의 음성 데이터에 대하여 계산한 power spectrum을 청각기의 주파수 반응도를 모사한 mel-scale 주파수 도메인에서 discrete cosine transform (DCT)를 취한 값이다. 일반적으로 음성인식에서 사용되는 MFCC 추출 과정은 하나의 frame에 대하여 12개의 DCT 결과 값과 전처리 과정을 거친 해당 frame 내 모든 음성 신호의 에너지 합, 총 13개의 coefficient로 구성된다. 이를 추출하기 위한 자세한 과정은 아래 동작 순서도와 같다. 각 블록에 대한 연산들은 [4]에 자세하게 기술되어 있다.

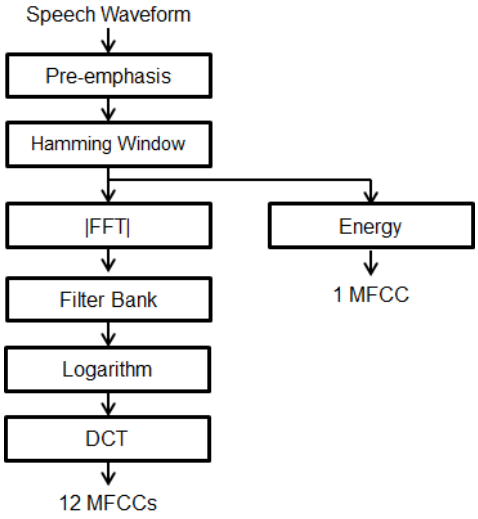

그림 1. MFCC 특징 추출 동작 순서도

그림 1 동작 순서도에서 다음과 같은 두 특성을 알 수 있다. 첫째로, MFCC 추출 과정의 중간 결과 값들은 dynamic range가 크다. FFT magnitude 및 energy를 계산하는 과정에는 해당 신호의 제곱 및 accumulate 연산이 사용되며, 그 결과 입력된 음성 신호 대비 큰 값이 출력된다. 반면 logarithm 함수를 거치면 입력 값 대비 작은 크기의 값이 출력된다. 따라서 서로 다른 두 신호를 logarithm 함수에 인가하였을 때 출력되는 두 값을 구분하기 위해서는 fractional part에 대한 resolution이 높아야 한다. 이러한 특성은 그림 2 예시에서 확인할 수 있다. 그림 2에서는 MFCC 추출 과정을 간소화한 연산에 대하여 fixed-point로 중간 결과 값을 표현하였다. 간소화한 연산은 입력 신호를 제곱하고 logarithm을 취하는 과정으로 구성된다. 이 때 입력 신호들은 상용 PCM 방식의 16-bit sample 마이크에서 얻은 값이다.

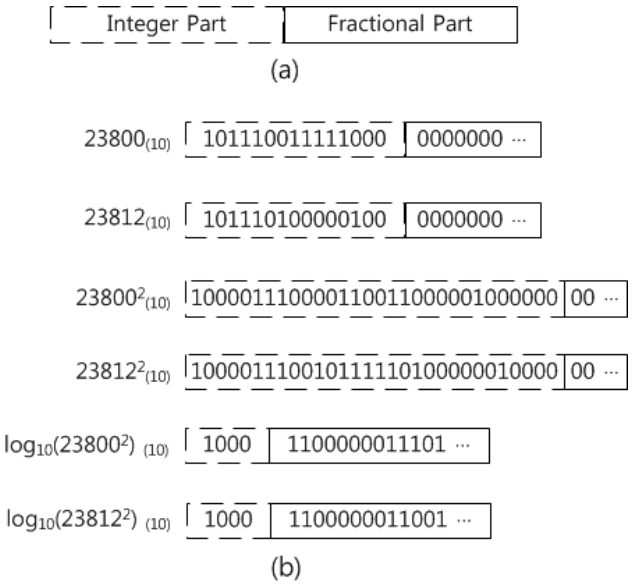

그림 2. 간소화된 MFCC 특징 추출 연산에 대한 예제 (a) Fixed-point 표현의 구성 요소 블록 표현 (b) MFCC 특징 추출 중간 결과에 대한 fixed-point 표현

그림 2에서 알 수 있듯이 fixed-point로 표현할 경우 integer part 및 fractional part의 비트 너비가 각 연산마다 비효율적으로 사용되는 것을 알 수 있다.

둘째로, MFCC 추출 과정에는 제곱근, logarithm과 같은 nonlinear 함수들이 존재한다. 상기 함수들은 독립 변수가 커질수록 양수인 1차 미분 값이 점차 줄어드는 특성을 가진다.

Floating-point는 동일한 비트 너비에 대하여 fixed-point보다 큰 dynamic range를 가진다. 따라서 MFCC 추출 과정을 앞의 특성에 부합하는 floating-point 연산으로 표현할 경우, 적은 비트 너비의 표현으로도 높은 음성인식 정확도를 달성할 수 있다.

2.2 Floating-point 표현

대표적인 floating-point 표현은 그림 3과 같이 부호(sign)를 표현하는 S, 지수부(exponent part)를 표현하

는 E, 정규화 된 실수부(fractional part)를 표현하는 F로 구성된다. 이 때 F는 0을 제외한 모든 값에 대하여 1.f′ 값으로 정규화 된다. 이 때 E는 2′s complement, F는 unsigned 수체계로 정의된다.

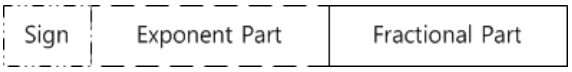

그림 3. Floating-point 표현의 구성 요소 블록 표현

해당 floating-point 표현의 값을 수식으로 나타내면 (1)과 같다. 나아가 제곱근 및 logarithm 연산의 경우, 각 함수의 특성에 따라서 수식 (2), (3)과 같이 연산을 변형할 수 있다.

$$(-1)^S \times F \times 2^E \quad (1)$$

$$\sqrt{F \times 2^E} = \sqrt{F} \times 2^{E/2} \quad (2)$$

$$\log_2(F \times 2^E) = \log_2(F) + E \quad (3)$$

변형된 연산을 사용할 경우 서로 다른 입력 {0,F,E_1}, {0,F,E_2}에 대하여 각 연산 별 동일한 $\sqrt{F}$ 및 $\log_2(F)$를 갖는다. 따라서 floating-point를 사용할 경우 제곱근, logarithm 연산에 필요한 look up table의 크기를 줄일 수 있다.

2.3 Floating-point 비트 너비 최적화 알고리즘

전력 소모를 최소화한 MFCC 추출 과정을 구현하기 위해서는 해당 음성인식 시스템의 합리적인 인식 정확도를 유지하면서 각 구성 블록 별 비트 너비를 최적화해야 한다. 이를 위하여 다음과 같이 알고리즘 A, B를 정의한다.

알고리즘 A는 그림 1 MFCC 추출의 모든 수체계를 double-precision floating-point로 정의한 음성인식 시스템 동작이다. 상기 알고리즘은 비트 너비 최적화 과정 중 사용되는 인식 정확도의 상한선을 결정한다.

알고리즘 B는 그림 1의 MFCC 추출 구성 요소 중 일부 블록에 대하여 floating-point의 fractional part 비트 너비를 변경한 후 음성인식을 수행한 동작이다. Fractional part의 비트 너비는 해당 수체계의 resolution을 결정하므로, 음성인식 정확도는 감소할 것이다.

앞서 정의한 알고리즘 A, B를 바탕으로 비트 너비 최적화 알고리즘을 그림 4와 같이 정의할 수 있다. 그림 4에 표시된 알고리즘 B는 pre-emphasis 블록의 비트 너비를 최적화하는 예시이다. 이 때 점선으로 표시된 블록은 floating-point의 fractional part 비트 너비가 변경되었음을 의미한다. 이를 제외한 나머지 블록은 double-precision floating-point 연산을 수행하였다.

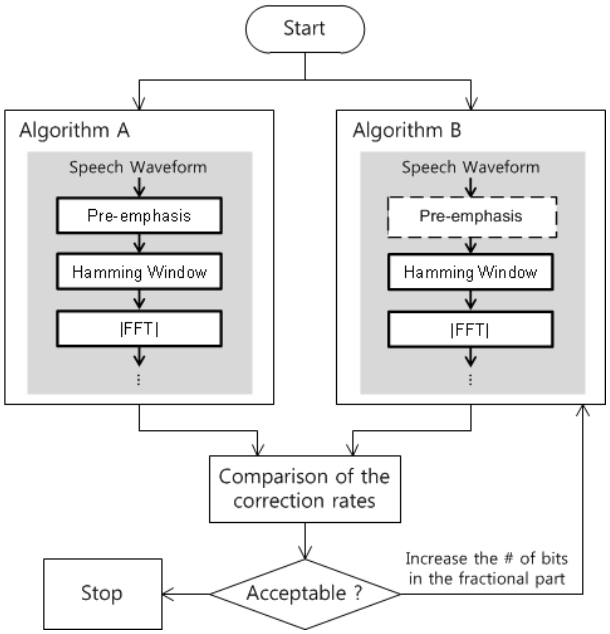

그림 4. Floating-point 비트 너비 최적화 알고리즘 동작 순서도

알고리즘 A에서 구한 인식 정확도의 상한선과 알고리즘 B에서 구한 인식 정확도를 비교하였을 때 차이가 threshold 값보다 클 경우 fractional part의 비트 너비를 1 증가시킨 후 알고리즘 B를 반복하고, threshold 값보다 작을 경우 해당 알고리즘을 멈춘다.

위 과정을 나머지 MFCC 추출 과정의 블록들에 대하여 반복한다. 모든 블록에 대하여 최적화된 fractional part의 비트 너비가 정해진 경우, 알고리즘 B에 대하여 앞서 구한 최적화된 비트 너비를 여러 블록에 적용한 후 그림 4의 최적화 알고리즘을 반복한다.

Floating-point의 exponent part 비트 너비는 각 블록 별 연산 결과 값이 overflow 및 underflow되지 않도록 정의한다.

Ⅲ. 구현

본 논문에서 제안한 알고리즘은 c언어로 구현된 음성인식기 simulator에 적용되었다. 음성인식기의 특징 추출 과정은 앞서 언급한 MFCC 특징 추출로, 음성 분류 과정은 Viterbi 알고리즘으로 구현하였다.

Isolated 단어 인식 과제에 대하여 제안하는 float-point 비트 너비 최적화 알고리즘을 적용하였을 경우 얻어진 floating-point MFCC 추출 과정의 비트 너비와 선행 연구에서 제시된 fixed-point MFCC 특징 추출 과정의 비트 너비[2]가 표 1에 비교되어 있다.

Function	Floating-point MFCC			Fixed-point MFCC [2]		
	Exp.	Frac.	Total	Int.	Frac.	Total
Speech	5	7	13	12	0	12
Pre-emp.	5	3	9	15	8	23
Hamming	3	2	6	15	9	24
$\|FFT\|^2$	7	4	12	42	8	50
$\sqrt{\|FFT\|^2}$	6	4	11	21	10	31
MF Bank	5	3	9	21	2	23
Log10	5	7	13	6	14	20
DCT	7	4	12	6	18	24
MFCC	7	4	12	6	14	20

표 1. 본 논문이 제시한 최적화된 floating-point MFCC 추출 과정의 비트 너비와 선행 연구 fixed-point MFCC 추출 과정의 비트 너비의 비교

표 1의 floating-point MFCC column에는 exponent part, fractional part의 비트 너비가 표현되어 있고 sign 비트는 생략되어 있다. 각 function 별 전체 비트 너비를 계산할 때에는 해당 sign 비트도 고려되었다. Fixed-point MFCC column에는 integer part 및 fractional part의 비트 너비와 이를 합한 전체 비트 너비가 적혀있다. Floating-point MFCC는 입력으로 16-bit sample 음성 신호를 사용하였음에도 불구하고 12-bit sample의 음성 신호가 사용된 fixed-point MFCC와 대등한 비트 너비의 음성 신호를 사용하여 음성인식 과정을 진행할 수 있었다. 대부분의 function에 대하여 본 논문에서 제시한 알고리즘을 적용한 방식의 비트 너비는 [2]보다 2배가량 작고, 최대 4.16배 작은 것을 확인하였다.

그림 5는 본 논문에서 제시한 최적화된 float-point 비트 너비의 MFCC 특징 추출 과정이 포함된 음성인식기 simulator가 정상적으로 테스트 단어를 인식하는 것을 출력한 화면이다. 상기 음성인식기는 double-precision floating-point 연산에 비해 4% 인식률이 낮은 93% 정확도를 달성하였다.

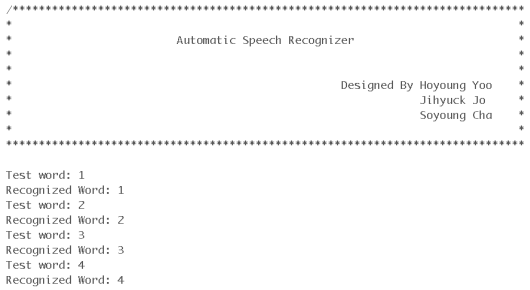

그림 5. 구현한 음성인식기 simulator의 동작 화면

Ⅳ. 결론 및 향후 연구 방향

본 논문에서는 음성인식 시스템 중 데이터 처리량이 많은 MFCC 추출 과정에 대하여 floating-point 연산의 타당성을 제시하고, 저전력 음성인식 시스템을 위한 floating-point 비트 너비 최적화 알고리즘을 제안하였다. 이를 적용한 음성인식기는 작은 인식 정확도 성능 저하를 수반하였지만, 최적화된 비트 너비가 선행 연구에서 제시한 시스템 블록들에 비하여 대부분 2배가량 작고, 최대 4.16배 작은 것을 확인하였다. 향후 해당 비트 너비의 floating-point 음성인식 시스템을 하드웨어로 구성하여 저전력 음성인식기를 개발할 예정이다.

Acknowledgement

이 논문은 정부(교육과학기술부)의 재원으로 (재)스마트 IT 융합 시스템 연구단(글로벌프론티어사업)의 지원을 받아 수행된 연구임 ((재)스마트 IT 융합시스템 연구단-20110031860)

참고문헌

[1] Wang, Jia-Ching, Jhing-Fa Wang, and Yu-Sheng Weng. "Chip design of MFCC extraction for speech recognition." INTEGRATION, the VLSI journal 32.1 (2002): 111-131.

[2] Ramos-Lara, Rafael, et al. "Real-time speaker verification system implemented on reconfigurable hardware." Journal of Signal Processing Systems 71.2 (2013): 89-103.

[3] www.sensoryinc.com/products/NLP-5x.html

[4] Zheng, Fang, Guoliang Zhang, and Zhanjiang Song. "Comparison of different implementations of MFCC." Journal of Computer Science and Technology 16.6 (2001): 582-589.