퍼셉트론 신경회로망을 사용한 유성음, 무성음, 묵음 구간의 검출 알고리즘

최재승*

Voiced-Unvoiced-Silence Detection Algorithm using Perceptron Neural Network

Jae-seung Choi*

요 약

본 논문에서는 다층 퍼셉트론 신경회로망을 사용하여 각 프레임에서의 유성음, 무성음, 그리고 묵음 구간을 검출하는 구간검출 알고리즘을 제안한다. 다층 퍼셉트론 신경회로망의 입력으로는 고속 푸리에변환에 의한 전력스펙트럼 및 고속 푸리에변환 계수가 사용되어 네트워크가 학습된다. 본 실험에서는 원 음성에 백색잡음이 중첩된 음성을 신경회로망에 입력함으로서 각 프레임에서의 유성음, 무성음, 묵음 구간의 검출성능 결과를 나타낸다. 본 실험에서는 신경회로망의 학습 데이터 및 평가 데이터가 다를 경우에도 이러한 음성 및 백색잡음에 대하여 92% 이상의 검출율을 구할 수 있었다.

ABSTRACT

This paper proposes a detection algorithm for each section which detects the voiced section, unvoiced section, and the silence section at each frame using a multi-layer perceptron neural network. First, a power spectrum and FFT (fast Fourier transform) coefficients obtained by FFT are used as the input to the neural network for each frame, then the neural network is trained using these power spectrum and FFT coefficients. In this experiment, the performance of the proposed algorithm for detection of the voiced section, unvoiced section, and silence section was evaluated based on the detection rates using various speeches, which are degraded by white noise and used as the input data of the neural network. In this experiment, the detection rates were 92% or more for such speech and white noise when training data and evaluation data were the different.

키워드

Detection algorithm, perceptron neural network, Fast Fourier transform, detection rate.

Ⅰ. 서 론

근년, 신경회로망을 사용한 음성인식을 실시하려고 하는 연구가 활발히 진행되며, 이러한 수법이 음성인식과 같은 일종의 애매함을 포함하는 문제의 해결에 유효함이 해결되어 왔다[1]. 이 중에서도 오차역전파 학습 알고리즘을 사용한 방법은 비교적 간단한 알고리즘임에도 불구하고, 패턴 인식에 있어서 상당히 강력한 학습 알고리즘이라는 것이 다수의 연구에 의하여 증명되고 있다[2]. 그러나 음성인식이 실용화되기 위해서는 아직 해결해야할 여러 가지 문제점이 남겨져 있다. 예를 들면, 마이크로폰에 있어서 주변으로부터 혼입되는 잡음의 영향에 의한 잡음제거의 문제, 불특정 화자, 음운 및 음절이 연속적으로 발성된 경우에 일어

* 신라대학교 전자공학과(jschoi@silla.ac.kr)
접수일자 : 2011. 02. 08 심사(수정)일자 : 2011. 03. 09 게재확정일자 : 2011. 04. 12

나는 조음결합 현상 등이다. 이 중에서도 잡음의 영향에 의한 잡음제거에 대해서는 각종 실용화의 경우를 고려하면 음성과 잡음의 혼재를 방지하는 것은 거의 불가능하며, 배경잡음을 제거하는 수법이 음성인식의 전처리로써 반드시 필요하다.

음성은 잡음에 의하여 비선형적으로 변화한다. 이 변화로부터 원래의 음성으로 복구가 가능하다면 음성인식의 전처리로써 사용한다든가 잡음을 제거하는 것이 가능할거라고 생각된다. 잡음이 중첩된 음성파형으로부터 직접 신경회로망 및 선형필터에 의하여 잡음을 제거하는 실험이 보고되고 있다[3]. 또한 잡음이 중첩된 환경 하에서 음성인식 방식으로서는 신경회로망에 의한 방법(Neural Network, NN)[4], 은닉 마르코프 모델 (Hidden Markov Model, HMM)[5]등의 방법들이 연구되고 있다.

잡음환경에서의 음성구간을 검출 할 경우에 아직도 많은 어려움이 존재하고 있다. 음성구간 검출의 기술은 음성신호의 유무를 판별하는 기술로서 잡음 환경하에서 음성인식시스템에 많이 적용되고 있으며, 일반적으로 끝점 검출알고리즘, short-term 에너지 변화를 이용하는 알고리즘, 주기성에 의한 피치검출 알고리즘, 신경회로망에 의한 검출 알고리즘 등이 연구되고 있다[6, 7]. 따라서 본 논문에서는 신경회로망의 학습 알고리즘을 사용하기 위하여 각 프레임에서 에너지변화를 이용하여 유성음, 무성음, 묵음 구간의 검출에 대한 알고리즘을 제안한다[8]. 본 실험에서는 신경회로망에 대해서 입력 신호대잡음비 SNRinput(Input Signal-to-Noise Ratio)을 Clean, 20 dB로 변경한 잡음이 중첩된 음성을 신경회로망에 입력함으로서 각 프레임에서의 유성음, 무성음, 묵음 구간의 검출 결과를 나타낸다.

II. 다층 퍼셉트론

현재, 음성의 분야에서도 음성의 규칙성에 있어서 문자로부터 음운에의 변환규칙의 발견에 3층의 신경회로망과 오차역전파학습 알고리즘(Back-Propagation Training Algorithm)[2]을 사용하여 성공함으로써 주목받아 오고 있다. 음성인식의 분야에 있어서도 신경회로망을 사용한 소수 카테고리의 인식이 실시되어, 종래 기술에 필적하는 인식율이 보고되기 시작하였다[9]. 이와 같이 현재 신경회로망은 단지 뇌구조와의 유사성뿐만 아니라 공학적으로도 유효하다는 가능성 때문에 상당히 주목받고 있으며, 특히 다층 퍼셉트론의 가중치를 오차역전파학습 알고리즘에서 학습시키는 수법이 널리 사용된다. 이러한 다층 퍼셉트론의 능력도 실험적으로 조사되고 있다[4].

본 논문에서 사용한 신경회로망은 중간층이 1층인 그림 1과 같은 다층 퍼셉트론[3]형의 계층형 네트워크를 사용하며, 네트워크의 유닛 간은 입력층으로부터 출력층으로 향하는 결합을 가진다.

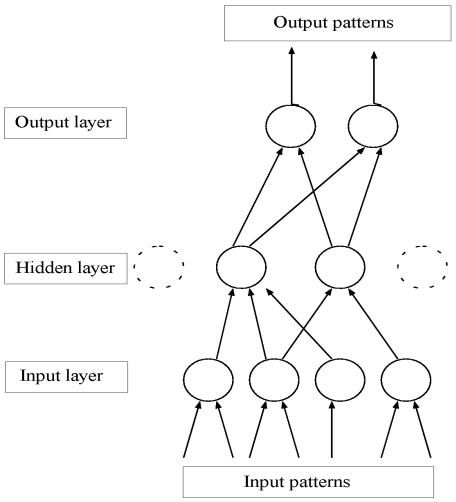

그림 1. 다층 퍼셉트론 신경회로망
Fig. 1 Multi-layer perceptron neural network

퍼셉트론형에서는 오차역전파학습 알고리즘을 사용하여 네트워크를 학습시키며, 이 알고리즘의 특징은 교사신호가 있는 학습에 있어서 출력층으로부터 입력층에 오차를 역전파시킴으로써 각 유닛에 대하여 최급강하법을 적용하며, 각 유닛에 비선형함수를 도입하여 입력으로부터 출력에의 사상을 가능하게 한다. 제 j 유닛의 출력을 o_j로 하여, 제 i 유닛으로부터 제 j 유닛에의 결합의 가중치를 w_{ji}로 하였을 때, 출력층과 중간층의 임의의 유닛은 고유의 가중치계수 w_{ji}에 의해서 모든 하위층의 유닛과 결합된다. 각 유닛의 입력 및 출력관계를 부여하는 비선형함수인 시그모이드

함수는 식 (1)과 같이 나타낸다. 여기에서 θ_i는 제 i 유닛의 문턱치(threshold)를 나타낸다.

$$f(x) = (\frac{2.0}{1.0 + \exp(-\sum_j w_{ji} o_j + \theta_i)}) - 1.0 \quad (1)$$

III. FFT 켑스트럼

켑스트럼 방법은 스펙트럼 대수의 척도에 의하여 구해지는 스펙트럼 포락에 의한 추정방법이다. 인간의 청각 기능이 외부의 자극에 대하여 대수적인 감도를 가지고 있기 때문에 켑스트럼 방법에서는 청각적으로 중요한 파라미터를 빠뜨리는 위험성이 적다고 판단된다. 본 논문에서는 고속 푸리에 변환(Fast Fourier Transform, FFT)에 의해서 구해지는 FFT에 의한 켑스트럼에 대해서 기술한다.

켑스트럼 방법은 켑스트럼에 창(window)를 씌움으로써 음원의 주기성에 대응하는 미세구조성분 g(t)를 제거하여, 스펙트럼 포락성분 v(t)의 단시간 영역성분만을 추출함으로써 평균화된 스펙트럼 성분을 구하는 방법이다. 일반적으로 캡스트럼 c(t)는 저역에 해당하는 quefrency 성분을 사용한다. 본 실험에서는 샘플링 주파수 8 kHz의 이산시간신호를 128샘플(16 ms)의 프레임으로 분리하여 각 프레임의 샘플값을 해밍창를 통과시킨 후에 켑스트럼 변환(FFT→log| |→IFFT)을 한다. 구해진 켑스트럼을 켑스트럼창에 통과시킴으로써 켑스트럼의 저역부의 10개의 켑스트럼 데이터를 구한다. 또한 입력 데이터에 대해서 FFT를 실시하여 FFT 전력스펙트럼을 구한다. 따라서 제안한 신경회로망 시스템에서는, 입력층의 유닛수는 10개의 켑스트럼 및 1개의 전력 스펙트럼의 총 11개를 신경회로망에의 입력으로 한다.

IV. 음성 및 잡음 데이터

본 실험에서 사용한 음성신호는 8 kHz의 샘플링 주파수를 가진 환경에서 녹음된 영어숫자로 구성된 Aurora2 데이터베이스(Database, DB)[10]를 사용하였다. Aurora2 DB의 모든 음성데이터는 ETSI (European Telecommunications Standards Institute)로부터 배포되었으며, 테스트 셋 A, B, C의 음성데이터로 구성되어 있다[11]. 본 실험에서는 Aurora2 DB의 테스트 셋 A, B, C 중에서 임의적으로 50문장을 선택하였으며, 20문장은 신경회로망의 학습 데이터로 사용하며 나머지 데이터는 평가용으로 사용하였다. 본 실험에서 사용한 잡음데이터는 컴퓨터에 의해서 작성된 가우스 백색잡음(white noise)의 배경잡음을 사용하여 평가하였다. 각 음성데이터마다 서로 다른 백색잡음을 중첩함으로써 SNR_{input}이 20 dB인 잡음이 중첩된 음성을 작성하였다. 또한, SNR_{input}으로서는 다음 식에서 나타내는 바와 같이 음성 S(n)과 잡음 N(n)의 전체에 해당하는 전력의 비율로서 정의되는 전역 SNR_{input}을 사용하였다.

$$SNR_{input} = 10 \cdot \log_{10}\left(\frac{\sum_{n=1}^{N} S(n)^2}{\sum_{n=1}^{N} N(n)^2}\right) \quad (2)$$

여기에서, N은 음성데이터의 샘플수이다.

V. 구간 검출시스템

본장에서는 3층 구조의 퍼셉트론형의 신경회로망에 FFT에 의한 전력스펙트럼 및 FFT에 의한 켑스트럼을 입력으로 하여 각 프레임에서 유성음, 무성음, 묵음에 대한 구간을 검출하는 것을 목적으로 하여 검출율을 높이는 실험에 대하여 기술한다. 본 실험에서의 음성 구간검출시스템의 평가방법으로는 각 프레임에서의 구간 검출율을 도입하였다. 이 검출율은 입력문장의 모든 프레임 수에 대하여 각 프레임에서 검출율이 정확하게 검출된 프레임수의 비율로 정의한다.

그림 2는 본 논문에서 제안하는 학습용 및 평가용의 문장이 동일한 경우의 검출시스템을 나타낸다. 본 실험에서는 잡음이 중첩된 음성신호를 128샘플(16 ms)의 프레임으로 분리한 후에 해밍창을 통과시킨다. Thresholding 블록에서, 해밍창을 통과한 잡음이 중첩된 음성신호는 각 프레임의 실효값 R_f가 문턱값

$R_m/3$보다 큰 경우에는 유성부(모음)으로 판별하도록 하며(즉, $R_f > R_m/3$인 경우), $R_m/5 \le R_f \le R_m/3$일 때에는 이 프레임은 무성부(자음)로 판별하며, $R_f < R_m/5$일 때에는 이 프레임은 묵음부로 각각 판별된다. 여기에서 R_f는 각 프레임에서 구해진 실효값을 나타낸다. 본 실험에서는 처음의 약 5프레임에서 각 문장의 평균 실효값 R_m을 실험적으로 구하였다. 유성부, 무성부, 묵음부로 각각 판별이 된 후에, 각 프레임의 음성신호 표본값으로부터 저역에 해당하는 10차의 FFT 켑스트럼을 구한다. 또한 입력 음성신호에 대해서 FFT를 실시하여 FFT 전력스펙트럼을 구한다. 여기에서 판별된 각 구간의 데이터에 대하여 -0.09~+0.09의 사이에 해당하는 값으로 정규화하여, 이 데이터들이 3층 구조의 신경회로망의 입력 데이터로 각각 부여되어 유성부, 무성부, 묵음부로 판별되도록 신경회로망이 학습된다. 그러나 본 논문에서 제안한 구간검출방법의 정확성을 위하여 다른 검출방법을 추가하는 등의 방법이 필요하다고 판단된다. 또한 본 구간검출 알고리즘의 효율성을 입증하기 위하여 pseudo code 등을 이용한 알고리즘 분석도 필요하다고 보며, 이러한 내용들은 향후의 연구과제로서 수행할 필요가 있다고 판단된다.

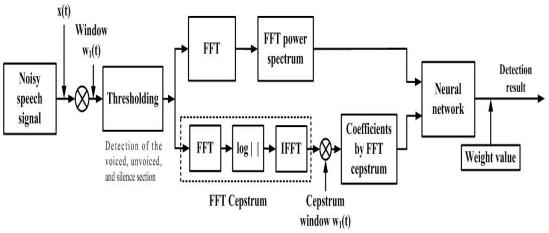

그림 2. 제안한 검출 시스템
Fig. 2 Proposed detection system

제안한 신경회로망 시스템에서는, 입력층의 유닛수는 10차의 FFT 켑스트럼 계수 및 1개의 FFT 전력스펙트럼의 총 11개를 신경회로망에의 입력으로 사용한다. 신경회로망에의 교사신호는 (T_1): 유성부 상태를 [-1.0, 0.0, 0.0], (T_2): 무성부 상태를 [0.0, -1.0, 0.0], (T_3): 묵음부 상태를 [0.0, 0.0, -1.0]으로 설정하여 유성부, 무성부, 묵음부를 각 프레임에서 인식하도록 신경회로망의 네트워크를 학습시킨다. 따라서 네트워크의 구성은 11개의 입력층 유닛, 17개의 중간층 유닛, 3개의 출력층 유닛으로 구성된 3층의 신경회로망에 입력함으로써, 각 출력신호는 학습신호와 일치한 정확한 값을 취하도록 네트워크를 학습시킨다. 본 실험에서는 학습계수 $\alpha = 0.1$, 가속도 계수 $\beta = 0.7$로 하였으며, 최대 학습횟수는 10,000회로 하였다.

VI. 검출시스템의 실험결과

본 실험에서 평가용으로 사용하는 음성은 Aurora2 데이터베이스의 테스트 셋 A, B, C로부터 잡음이 중첩된 음성 데이터들이 임의적으로 선택되었으며, 잡음 데이터는 학습 시에 사용한 동일한 잡음인 백색잡음이 선택되었다.

본 실험에서는 신경회로망의 학습을 통해 구해진 가중치의 출력 결합계수를 저장한 후, 학습에 사용한 잡음이 중첩된 음성신호 및 학습에 사용하지 않은 잡음이 중첩된 음성신호의 FFT 전력스펙트럼 및 FFT 켑스트럼계수를 각각 신경회로망의 입력으로 사용하여 교사신호 T_1, T_2, T_3의 목표값과 비교하여 각 프레임에서 검출율을 구한다. 표 1의 (a), (b), (c)는 신경회로망의 학습데이터로서 음성(M1)만을 사용하여 학습을 실시하여, 학습 시에 사용한 동일한 음성신호(M1) 및 학습 시와 다른 음성신호(M2, F1)를 신경회로망의 입력으로 사용한 경우의 구간 검출율에 대한 실험결과이다. 표 1(a)는 $R_f > R_m/3$, 즉 유성부가 신경회로망에 입력된 경우의 구간 검출율을, 표 1(b)는 $R_m/5 \le R_f \le R_m/3$, 즉 무성부가 신경회로망에 입력된 경우의 구간 검출율을, 표 1(c)는 $R_f < R_m/5$, 즉 묵음이 신경회로망에 입력된 경우의 구간 검출율을 각각 나타낸다. 표 1의 (a), (b), (c)의 결과로부터 학습데이터 및 평가데이터가 다른 경우에 대해서도 구간 검출율은 최대 92% 이상인 것을 알 수 있다. 본 실험에서는 신경회로망의 학습데이터 및 평가데이터로서 남성화자인 M1, M2 및 여성화자인 F1, F2를 사용하였다.

표 1. 음성 학습데이터에 대한 검출율(%)
Table 1. Detection rates for the speech data.

(a) 유성부의 경우(In the case of voiced section)

학습 데이터	평가 데이터	구간 검출율(%)		
		유성음	무성음	묵음
M1	M1	92.1%	7.9%	0.0%
	M2	89.3%	9.9%	0.8%
	F1	87.6%	11.2%	1.2%

(b) 무성부의 경우(In the case of unvoiced section)

학습 데이터	평가 데이터	구간 검출율(%)		
		유성음	무성음	묵음
M1	M1	9.4%	90.0%	0.6%
	M2	11.6%	87.1%	1.3%
	F1	12.9%	85.2%	1.9%

(c) 묵음부의 경우(In the case of silence section)

학습 데이터	평가 데이터	구간 검출율(%)		
		유성음	무성음	묵음
M1	M1	0.8%	11.0%	88.2%
	M2	1.7%	12.9%	85.4%
	F1	2.1%	14.8%	83.1%

표 2의 (a), (b), (c)는 신경회로망의 학습데이터로서 원 음성(F1)에 백색잡음을 중첩시킨 SNR_{input}=20 *dB*에 대하여 잡음이 중첩된 음성신호를 사용하여 학습을 실시하여, 학습 시에 사용한 동일한 음성신호(F1) 및 백색잡음, 그리고 학습 시와 다른 음성신호(F2, M2) 및 백색잡음을 신경회로망의 입력으로 사용한 경우에 대해서, 유성부, 무성부, 묵음부가 각각 신경회로망에 입력된 경우의 구간 검출율에 대한 실험결과를 나타낸다. 표 2의 (a), (b), (c)의 결과로부터 학습데이터 및 평가데이터가 동일한 경우의 검출율은 최대 86% 이상인 것을 알 수 있다.

지금까지 기술한 표의 결과로부터 알 수 있듯이, SNR_{input}이 20 *dB*인 경우에 대해서 FFT 켑스트럼 및 NN를 사용하여 각 프레임에서 유성음, 무성음, 묵음의 구간 검출이 양호하게 인식되는 것을 알 수 있다.

표 2. 잡음이 중첩된 음성의 학습데이터에 대한 검출율(%)
Table 2. Detection rates for the noisy speech data.

(a) 유성부의 경우(In the case of voiced section)

학습 데이터	평가 데이터	구간 검출율(%)		
		유성음	무성음	묵음
F1	F1	86.3%	13.2%	0.5%
	F2	82.5%	16.2%	1.3%
	M2	81.6%	15.9%	2.5%

(b) 무성부의 경우(In the case of unvoiced section)

학습 데이터	평가 데이터	구간 검출율(%)		
		유성음	무성음	묵음
F1	F1	15.5%	83.8%	0.7%
	F2	18.0%	80.2%	1.8%
	M2	18.4%	79.3%	2.3%

(c) 묵음부의 경우(In the case of silence section)

학습 데이터	평가 데이터	구간 검출율(%)		
		유성음	무성음	묵음
F1	F1	0.9%	17.6%	81.5%
	F2	1.9%	20.0%	78.1%
	M2	2.3%	20.7%	77.0%

VII. 결 론

본 논문에서는 다층 퍼셉트론 신경회로망을 사용하여 유성부, 무성부, 그리고 묵음부에 대한 각 프레임에서의 구간 검출에 대한 검출 시스템을 제안하였다. 제안한 시스템은 SNR_{input}가 20 *dB*인 경우에 대하여 유성부, 무성부, 묵음부를 검출한 후에 오차역전파 알고리즘에 의한 3층 구조의 퍼셉트론 신경회로망을 사용하여 학습하였으며, 신경회로망의 입력으로는 FFT 켑스트럼에 의한 10개의 켑스트럼 및 1개의 전력 스펙트럼의 총 11개를 신경회로망에의 입력으로 하였다. 신경회로망의 학습결과로부터 SNR_{input}=20 *dB*에 대하여 학습데이터 및 평가데이터가 동일한 경우, 원 음성에 백색잡음을 중첩시켰을 때의 검출율은 최대 86%

이상인 것을 알 수 있었다.

이상으로 본 논문에서 제안한 검출 시스템을 다층 퍼셉트론 신경회로망을 사용하여 본 알고리즘이 백색 잡음에 대해서 유효하다는 것을 알 수 있었다. 향후의 연구과제로는 자동차잡음 등의 다양한 배경잡음을 사용하여 본 논문에서 제안한 검출시스템의 향상 기법을 검토할 필요가 있다고 본다. 또한 SNR_{input}을 다양하게 검토하여야 할 필요가 있다고 본다.

참고 문헌

[1] L. Tan, P.C. Ching, L.W. Chan, “Recurrent neural networks for speech modeling and speech recognition”, International Conference on Acoustics, Speech, and Signal Processing, vol.5, pp. 3319 - 3322, 1995.

[2] D.E. Rumelhart, G.E. Hinton, and R. J. Williams, "Learning representations by back-propagation errors", Nature, vol.323, pp. 533-536, 1986.

[3] T.T. Le, J.S. Mason and T. Kitamura, "Characteristics of multi-layer perceptron models in enhancing degraded speech", Proc. ICSLP-94, pp. 1611-1614, 1994.

[4] R.P. Lippmann, "An Introduction to Computing with Neural Nets", IEEE ASSP Magazine, vol.4, no.2, pp. 4-22, April 1987.

[5] T. Hirahara and H. Iwamida, "Auditory spectrograms in HMM phoneme recognition", Proc. Int. Conf. Spoken Lang. Process., ICSLP-90, pp. 1381-1384, 1990.

[6] K. Yamamoto, F. Jabloun, K. Reinhard, A. Kawamura, "Robust Endpoint Detection for Speech Recognition Based on Discriminative Feature Extraction", IEEE International Conference on Acoustics, Speech and Signal Processing, vol.1, pp. I.805-I.808, 2006.

[7] W. Kun-Ching, T. Yi-Hsing, "Voice Activity Detection Algorithm with Low Signal-to-Noise Ratios Based on Spectrum Entropy", Second International Symposium on Universal Communication, pp.423-428, 2008.

[8] 최재승, “다층 퍼셉트론 신경회로망을 사용한 구간 검출 알고리즘”, 한국해양정보통신학회 추계학술대회 논문집, 14권, 2호, pp. 274-277, 2010.

[9] H. Leung and V. Zue, "Some phonetic recognition experiments using artificial neural nets", ICASSP 88, pp. 422-425, 1988.

[10] H. Hirsch and D. Pearce, “The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions,” in Proc. ISCA ITRW ASR2000 on Automatic Speech Recognition: Challenges for the Next Millennium, Paris, France, 2000.

[11] R.G. Leonard, “A database for speaker independent digit recognition,” IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.328-331, Mar 1984.

저자 소개

최재승(Jae-seung Choi)

1989년 조선대학교 전자공학과 공학사

1995년 일본 오사카시립대학 전자정보공학부 공학석사

1999년 일본 오사카시립대학 전자정보공학부 공학박사

2000년~2001년 일본 마쯔시타 전기산업주식회사(현, 파나소닉 주식회사) AVC사 연구원

2002년~2007 경북대 디지털기술연구소 책임연구원

2007년~현재 신라대학교 전자공학과 교수

※ 관심분야 : 음성신호처리, 신경회로망, 적응필터와 잡음제거, 디지털 TV 및 멀티미디어 등