무선 통신망에서 음성인식률 개선을 위한 보상기법 연구

서 진 호, 박 호 종
광운대학교 전자공학과

Compensation Method for Improvement of Speech Recognition in Wireless Communication Network

Jin-Ho Seo, Ho-Chong Park
Dept. of Electronics Engineering, Kwangwoon University
gopassseo@kw.ac.kr, hcpark@kw.ac.kr

요약

이동통신 기술의 발전으로 이동통신 사용이 폭발적으로 증가하였고 그에 따라 이동통신망을 이용한 많은 서비스가 제공되고 있다. 이동통신망에서의 음성 인식 서비스에서 음성 인식기에 입력되는 음성신호는 통신망을 통해 음성 압축기를 거치게 되고 이에 음성신호가 왜곡되어 인식기의 인식성능이 저하 된다. 본 논문에서는 무선통신 환경에서 음성인식기의 성능을 개선하기 위한 보상 방법을 제안 한다. 기존의 제안된 방법은 음성 데이터에 의존하는 방법을 사용하나 본 논문에서는 음성 데이터와는 독립적 방법인 음성 압축기에 의해 손상된 입력 신호의 스펙트럼 보상방법과 Cepstrum 보정방법을 통해 인식률을 향상시키는 방법을 제안한다. 즉, 음성 압축기에 의하여 왜곡된 스펙트럼을 단계적 방법으로 보상하고 그를 토대로 왜곡된 신호에서 만들어진 Cepstrum을 보정하여 음성 인식기의 성능을 향상시키는 방법을 연구하였으며, 그 결과 손상된 음성신호의 인식률 64.88%에 대하여, 본 논문에서 제안하는 보상 방법을 적용한 음성신호의 인식률은 79.73%로서 14.85%가 향상된 결과를 얻을 수 있었다.

1. 서론

이동통신 환경에서의 음성인식분야의 접목은 이제껏 음성인식 분야에서는 고려하지 않았던 새로운 문제점을 가지고 있다. 그것은 인식 시스템을 구성할 때 모든 요소들이 한 시스템 안에 존재 할 수 없다는 것이다. 즉 하나의 시스템이 역할에 따라 구분되어 각각의 요소들이 자신의 맡은 처리만 할 뿐 각 요소들 간의 긴밀성이 없어져 버린다는 것이다. 긴밀성이 없어진다는 것은 구분되어진 시스템 사이에 어떤 일을 하는 부가적인 요소들이 끼어들 수 있다는 이야기이다. 이는 인식시스템이 어떠한 말을 인식한다는 전체적인 입장에서 볼 때 중간에 끼어든 다른 요소들이 인식과정에 어떠한 영향을 줄 수 있다는 이야기이다. 간단하면서 본 논문에서도 다룰 예를 들어본다. "이동통신 단말기에서 전화 서비스 업체에 전화를 걸어 서비스를 받으려고 한다. 전화 서비스 업체는 음성인식 시스템을 도입하여 서비스를 하고 있다. 그러므로 전화를 건 사용자는 서비스를 받기위하여 음성으로 명령을 내린다." 이 상황을 생각해보면 명령을 내리기위한 음성입력은 이동통신 단말기에서 받게 되며 그 입력을 받아 명령을 인식하여 서비스를 해주는 것은 전화 서비스 업체이다. 입력과 인식시스템이 분리가 되어 있고 그 사이에 무선 통신망 채널과 음성압축기라는 요소가 들어가 있다. 그림 1에서 그 구조를 보여주고 있다. 이 구조의 인식 시스템은 위에서 언급한 두 가지 요소에 민감하게 작용하는 것을 실험을 통해 알 수 있었다. 본 논문에서는 이동통신망 구조의 음성인식에서 음성압축기란 요소에 대한 음성인식률의 변화를 측정하며 이 요소로 인하여 음성인식률이 저하되는 것을 막기 위해 음성압축기의 성격을 분석하여 그에 맞는 보상방법을 제안 한다.

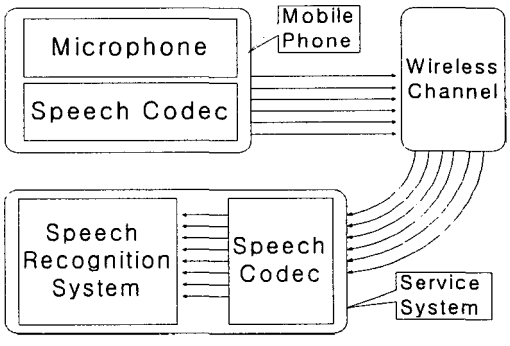

그림 1. 이동통신망에서의 음성인식 시스템 구조

2. 이동통신망에서의 음성인식 시스템

2.1 음성인식 시스템의 구조

그림 2(a)는 기본적인 음성인식 과정을 나타낸다. 서비스에 사용될 언어들을 미리 훈련시켜 인식 시스템에 저장하며 이것을 토대로 실제 입력되는 음성을 인식하게 되는 것이다. 이 시스템을 이동통신환경에 적용할 경우 두 가지 구조가 나올 수 있다. 그림 2(b)와 2(c)에서 그 구조를 살펴볼 수 있다. 음성압축기로부터 복원된 신호에서 특징을 추출하는 구조와 음성 비트 스트림에서 특징을 추출하는 구조이다. [2]의 연구에서 복원된 신호에서 특징을 추출하는 것 보다 비트 스트림에서 특징을 추출하는 것이 더욱 성능이 좋으며 채널에러가 심할 때는 더욱 성능이 차이가 난다는 것을 증명하였다. 그러나 이 방식을 쓰려면 이동통신 교환기 내부에 음성인식 모듈이 들어가야 하며 이는 통신시스템을 전부를 고쳐야 하는 결과를 초래하게 되어 시스템 적용에 어려움이 따른다. 따라서 보다 간단하게 음성 인식을 구현하려면 음성압축기에서 복원된 신호를 가지고 특징을 추출하는 방식을 일반적으로 사용한다.

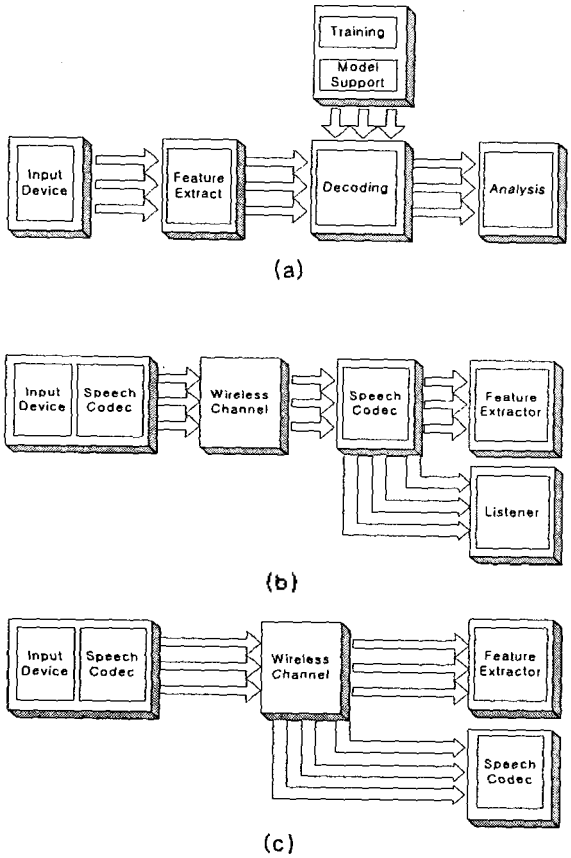

그림 2. 음성인식 과정 및 시스템 구조
(a)기본적인 음성인식 과정
(b)복원된 신호로부터 특징을 추출하여 인식하는 구조
(c)음성 비트 스트림으로부터 추출하여 인식하는 구조

2.2 음성압축

본 논문에서 사용된 음성압축기 IS-127 EVRC[4], ITU G.729, IS-96 QCELP 이다. 이들 중 IS-127 EVRC, IS-96 QCELP은 실제로 이동통신에서 사용되고 있는 음성압축기이다. 3개의 음성압축기는 사람의 음성신호 발생 모델을 기초로 하여 음성신호에서 특징을 추출해서 그 정보를 보내는 형식을 채택하고 있다. 보내는 정보는 LPC(Linear Prediction Coefficient)와 Pitch정보, 그리고 모델링 하지 못한 부분을 보정해주는 Fixed CodeBook에 대한 정보이다. 이와 같이 이들 압축기는 CELP라는 같은 개념으로 개발된 음성압축기 이지만 각기 다른 특징들을 가지고 있다. EVRC와 G.729는 Algebraic CodeBook을 사용하며 QCELP는 Random Circular CodeBook을 사용한다. 특히 EVRC는 입력단에 잡음 제거기를 가지고 있어 잡음을 줄일 뿐 아니라, RCELP라는 기술을 쓰고 있다. 이는 예측된 Pitch 구조를 기준으로 현재 입력 Pitch신호를 시간적으로 Warping을 시키는 기술로서 Pitch 잔여신호를 최소화하여 Fixed CodeBook이 표현해야하는 오차를 줄여 음질의 향상을 가져온다. 그림 2.에서는 원 음성신호와 원 음성신호를 각각의 음성압축기에 넣어 얻은 신호의 스펙트럼을 비교하여 나타낸 것이다.

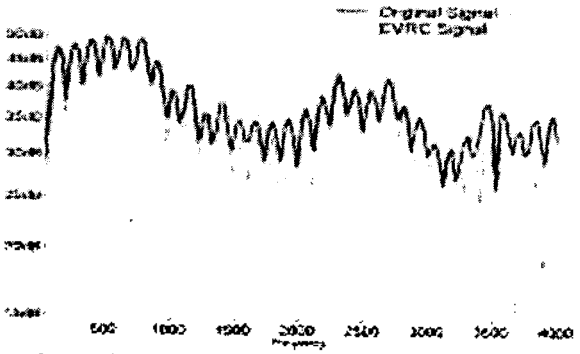

그림 3. 원 음성신호와 EVRC의 스펙트럼 비교

3. 제안하는 알고리즘

그림 3에서 보듯이 음성압축기는 음성신호를 왜곡 시키며, 실험에 의하면 이 신호를 그대로 음성인식 시스템에 사용하면 왜곡되지 않은 신호를 사용하였을 때 보다 인식률이19.66% 감소한다. 인식률이 떨어지는 것은 음성압축기에 의한 스펙트럼 에너지의 감소라고도 볼 수 있으며 이점을 착안하여 왜곡된 음성신호의 스펙트럼을 보상하는 알고리즘과 Cepstrum을 보정하는 방법을 연구하였다. 선행연구에서는 특징 추출 과정에 스펙트럼 보상 모듈을 삽입하여 왜곡된 음성신호를 보상하였으나 본 논문에서는 스펙트럼 보상법을 원 음성신호의 가상 타겟 신호를 만드는데 사용하고, 이것을 토대로 Damaged Cepstrum을 보정하는데 사용한다. 그림 4는 두 방법의 블록 다이어그램을 나타내고 있다. 왜곡된 음성신호의 스펙트럼 보상만으로는 더 이상 만족할 수 있는 결과를 얻어내기는 힘들다는 것을 선행연구에서 알게 되었다.

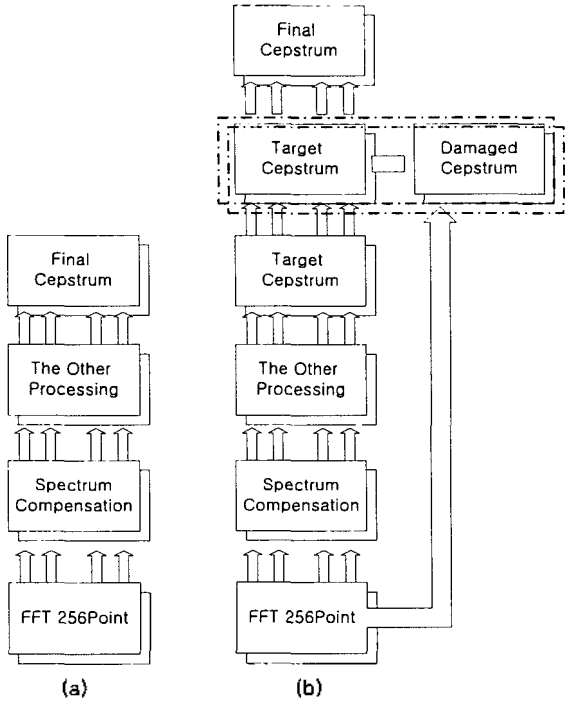

그림 4.스펙트럼 보상방법과 Cepstrum 보정법
(a) 선행연구에서 연구하였던 스펙트럼 보상방법
(b) Target을 두고 Cepstrum을 보상하는 Cepstrum 보정법

좀 더 세밀한 보정방법이 필요하였으며 그 방안으로 제안하는 것이 Cepstrum 보정방법이다. 먼저 Cepstrum 보정 전 단계인 스펙트럼 보상 방법을 간단히 살펴보자. 스펙트럼의 보상방법은 단계적으로 나뉘어져 있다. 첫 번째는 Sorting[1]에 의한 가중치 함수 적용 방법이다. 그림 3에서 볼 수 있듯이 왜곡된 음성신호에서 크기가 큰 음성신호는 원 음성신호를 잘 따라가고 있기에 가중치 함수를 적용할 필요가 없다 그러나 작은 신호들과 고주파수 영역의 신호들은 왜곡이 많이 되고 있기 때문에 스펙트럼을 Sorting 하여 크기가 작은 순서대로 나열하고 크기에 따른 가중함수를 적용한다. 수식 1에 나와 있는 것과 같이 프레임의 스펙트럼을 구한 뒤 평균값을 구하고 그 평균값과 가장 비슷한 스펙트럼을 찾아 가중치 함수를 적용 하는 범위를 정하는데 쓴다. 이 때 가중함수의 적용범위를 전체 스펙트럼의 평균값을 사용하여 구하는 이유는 각각 처리되는 프레임의 스펙트럼 성격을 반영하여 유동적 처리를 하기 위해서이다.

$$\begin{cases} Frame_i = \{sp_1, sp_2, sp_3 \ldots sp_{128}\} \\ i = Frame\ Number,\ sp_{1\sim128} = spectrum(256PointFFT) \end{cases}$$

$$\begin{cases} Mean = \dfrac{\left(\sum Frame_i = \{sp_1, sp_2, sp_3 \ldots sp_{128}\}\right)}{128} \\ Sp_{Mean} = FindIndex[sp_i \simeq Mean_{sp}] \end{cases}$$

$$\begin{cases} Weight = 1 & Ranking = \{Index[the\ other]\} \\ Weight\ Func. & Ranking = \{Index[Min_{sp} \sim Sp_{Mean}]\} \end{cases}$$

수식 1. Sorting 에 의한 가중함수 적용

두 번째는 고주파수 영역의 예외적 보상처리 이다. 만약 수식 2에서처럼 고주파영역의 스펙트럼이 가중치를 1.0을 받게 될 때 고주파수 영역의 스펙트럼만을 빼내어 예외 가중치함수를 적용시켜준다.

$$\begin{cases} Weight = 1 & Ranking = Index\begin{pmatrix} HighFrequencyBand_{sp_1} \\ HighFrequencyBand_{sp_2} \\ the\ other\ Band_{sp} \end{pmatrix} \\ Weight\ Func. & Ranking = \{Index[Min_{sp} \sim Sp_{Mean}]\} \end{cases}$$

$$\begin{cases} Weight = 1 & Ranking = \{Index[the\ other]\} \\ Weight\ Func. & Ranking = \{Index[Min_{sp} \sim Sp_{Mean}]\} \end{cases}$$

$$ExceptionWeightFunction = Index\begin{pmatrix} HighFreq.Band_{sp_1} \\ HighFreq.Band_{sp_2} \\ HighFreq.Band_{sp_3} \end{pmatrix}$$

수식 2. 고주파영역의 스펙트럼 예외처리

세 번째 단계는 Time Smoothing 기법이다. 특징 추출의 해상도를 높이기 위해 프레임간 오버랩 처리를 한다. 오버랩 되는 프레임간의 Correlation은 상당히 높기 때문에 가중치 함수를 적용할 때 오버랩 되었던 프레임의 가중치 정보를 이용하면 좋은 결과를 얻을 수 있다. 그림 5는 이 방법을 도시하고 있다. 현재 프레임이 N 프레임이며 가중치 함수를 적용하려고 하고 있다. 이 때 바로 이전 프레임인 N-1 프레임의 가중치 함수 적용 정보를 가져와서 현재 프레임의 가중치 함수정보와 비교해 본다. 바로 이전프레임인 N-1프레임에서는 가중치를 받는 스펙트럼이 현재 프레임에서는 간소한 차이로 가중치 함수를 적용받지 못한다. 이 때 현재 프레임이 받기로 결정된 가중치 범위에서 더욱 확장하여 과거의 정보역시 가중치를 받도록 한다. 물론 추가적으로 가중치를 받는 스펙트럼의 가중치 함수는 시간에 따른 비율계산으로 산정하게 되며 시간의 차이가 많이 날수록 비율은 더 낮아진다. 물론 지난 프레임에 주는 가중치 함수는 시간에 따라 낮은 비율로 적용이 된다.

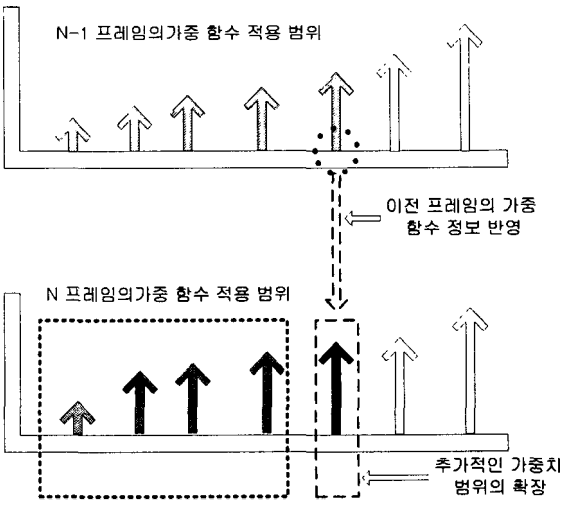

그림 5. Time Smoothing 기법

이와 같은 방법으로 스펙트럼을 보상한 뒤 최종적인 Cepstrum 계수를 만들어 낸다. 이렇게 만들어진 Cepstrum 계수와 원 음성신호로 만든 Cepstrum 계수의 Cepstrum Distance[3]가 원 음성신호로 만든 Cepstrum 계수와 왜곡된 신호를 보상하지 않고 만든 Cepstrum의 Cepstrum Distance 보다 더 작다. Cepstrum Distance 측정 결과 포함 즉 왜곡된 음성신호의 스펙트럼을 보상하여 만든 Cepstrum은 왜곡된 음성신호를 보상하지 않고 만든 Cepstrum 보다 원 음성신호의 Cepstrum에 더 가깝다. 이러한 사실을 이용하여 Cepstrum Distance를 좀 더 줄여 원 음성신호의 Cepstrum에 근접하려고 하였다. 왜곡된 신호를 보상하여 만든 Cepstrum을 Target 명명한다. 위와 같은 방법으로 구한 Target과 왜곡된 음성신호로 만든 Cepstrum 계수간의 계수별 차이를 구한다. 이 때 Sign과 Distance값이 나오며 이 때 Sign 값은 Distance의 방향을 나타낸다. 이렇게 구한 두개의 값을 가지고 Target(Cepstrum 계수)에 적용한다. 적용 방법은 각각의 Target이 가지고 있는 Sign의 방향으로 Distance를 조씩 늘려 원 음성신호의 Cepstrum과 차이를 줄인다.

4. 실험 및 결과

인식 실험은 HTK 3.1버전을 사용하여 고립단어 인식을 하였으며 단어 단위의 인식 실험을 하였다. 실험에 쓰인 데이터는 현재 이동통신에서 서비스되는 음성 퀄리티와 맞추기 위해서 8khz의 샘플링을 갖는 음성 데이터를 사용하였다. 음성 데이터는 SITEC에서 제작한 "클린 스피치 PBW452단어 DB"이며 훈련 데이터는 클린 스피치 30명분, 개인당 452단어씩 총 13560단어이며 테스트 데이터는 8명분, 3616단어이다. 실험은 두 개의 음성인식 모델을 만들었다. 하나는 원 음성신호만으로 구성된 훈련데이터로 만들었으며 다른 하나는 원 음성신호와 3개의 음성압축기 (IS-127 EVRC, ITU G.729, IS-96 QCELP)를 거친 훈련데이터로 만들었다. 원 음성신호로 만들었던 인식 모델에는 원 음성신호 테스트 데이터와 3개의 음성압축기를 거친 테스트 데이터, 또 음성압축기를 거친 테스트 데이터를 제안된 방법으로 보상하여 만든 데이터를 실험 하였고 다른 한 개의 인식모델에는 원 음성신호 데이터와 3개의 음성압축기를 통과한 음성데이터만 가지고 실험을 하였다.]

표 2.에서EVRC, QCELP 음성압축기에 대한 실험결과를 보면 성능의 뚜렷한 차이를 볼 수 있는데 이는 본 논문에서 제안하는 방법을 적용한 3개의 음성압축기 중에서 가장 민감하다고 할 수 있다. 즉 CELP구조의 음성압축기 중에서 잡음제거기, RCELP, Random Circular CodeBook 등 이런 특징적 방법을 쓰는 음성압축기가 본 논문의 제안방법이 잘 맞는다고 할 수 있다.

표 2. 실험 결과

훈련조건 / 테스트 데이터	클린 스피치로 Model 생성		클린 스피치와 왜곡신호로 Model 생성
	제안방법 적용 전	제안방법 적용 후	기존의 제안된 방법의 결과
Origianl Signal	84.29%	×	84.54%
EVRC	64.88%	79.73	82.66%
G.729	74.06%	78.35%	85.59%
Qcelp8k	66.23%	77.13%	83.05%

5. 결론

음성압축기를 거쳐 왜곡된 음성신호로 음성인식을 할 경우의 인식률저하를 막을 방법이 필요하였고 그 방법들 중 왜곡된 음성신호의 정보를 인식모델에 적용함으로써 인식모델을 강인하게 하는 데이터 의존적인 방법이 제일 좋은 결과를 보여주고 있다. 왜곡된 음성신호의 정확한 정보를 음성인식기에 알려준다는 의미에서 어쩌면 이 보다 더 나은 방법이 없을 것 같기도 하다. 하지만 다른 분야들의 기술 발전에 힘입어 인식 분야와 통합서비스를 많이 시도하려고 하는 이 시점에서 기존의 데이터에 의존한 방법에만 기대고 있는 것은 큰 문제가 아닐 수 없다. 따라서 다른 분야와 접목이 될 때 접목되는 분야의 특징적 상황과 기술을 이용해 문제점을 해결해 나간다면 더 좋은 성과를 거둘 수 있다고 생각된다. 그런 관점에서 볼 때 본 논문에서의 연구는 이동통신기술과 음성인식분야의 접목 시 발생되는 문제점의 새로운 해결 방안제시 라는 점에서 큰 의미를 가진다고 할 수 있다. 앞으로 두 기술 분야의 체계적인 분석과 긴밀한 기술의 조합이 필요하며 이것이 이루어졌을 때 현재 나와 있는 인식 성능보다 더 나은 결과를 얻을 수 있을 것 이다.

참고문헌

[1] 한상욱, 박호종, "이동 통신망에서의 음성인식률 향상 한국음향학회 추계 학술대회, 고려대학교, 2003년

[2] Alexis Bernard and Abeer Alwan, "Low-Bitrate Distributed Speech Recognition for Packet-Basedand Wireless Communication", IEEE Trans. on Speech and Audio Processing, Vol. 10, NO.8, November, 2002.

[3] Lawrence Rabiner, Biing-Hwang Juang Fundamentals of SpeechRecognition, Prentice-Hall International, inc. pp.149-193, 1993.

[4] TIA/EIA/IS-127 "Enhanced variable rate codec, speech service option 3 for wideband spectrum digital systems,", 1996.