한국음향학회지 제23권 제8호 pp. 597~601 (2004)

CDMA이동통신환경에서의 음성인식을 위한 왜곡음성신호 거부방법

Distorted Speech Rejection For Automatic Speech Recognition under CDMA Wireless Communication

장 준 혁*, 김 남 수**
(Joon-Hyuk Chang*, Nam Soo Kim**)

*캘리포니아 주립대학, 산타바바라, **서울대학교 전기컴퓨터공학부

(접수일자: 2004년 9월 9일; 수정일자: 2004년 10월 14일; 채택일자: 2004년 11월 11일)

본 논문에서는 CDMA이동통신 환경에서의 음성인식을 위한 왜곡음성신호의 전처리-거부방법을 소개한다. 먼저, CDMA이동통신 채널에서의 왜곡된 음성신호를 분석하고 분석된 매커니즘을 바탕으로 채널에 의해 왜곡된 음성신호를 음성의 준주기성을 바탕으로 하여 거부하는 알고리즘을 제안한다. 실험을 통해 제안된 전처리-거부방법이 적은 계산량을 가지고 음성인식에 적용되어 효과적으로 CDMA에 환경에서 채널왜곡된 음성신호를 거부할 수 있음을 알 수 있었다.

핵심용어: 전처리-거부방법, 자동음성인식, 준정상성, CDMA
투고분야: 음성처리 분야 (2.5)

This paper introduces a pre-rejection technique for wireless channel distorted speech with application to automatic speech recognition (ASR). Based on analysis of distorted speech signals over a wireless communication channel, we propose a method to reject the channel distorted speech with a small computational load. From a number of simulation results, we can discover that the pre-rejection algorithm enhances the robustness of speech recognition operation.

Keywords: Pre-rejection, Automatic Speech recognition, Quasi-stationarity, CDMA
ASK subject classification: Speech signal processing (2.5)

I. 서 론

최근에 자동음성인식 (automatic speech recognition, ASR) 기술은 실제적으로 많은 발전을 이루어 왔고, 편리하고 다양한 서비스를 가능하게 하였다. 음성인식 관련 시스템은 일반적으로 이동통신망을 통해 음성인식 서버와 이동장비간을 전화망 (유/무선, 이동통신)으로 연결해서 사용하는 것이 일반적이며, 이런 경우가 보다 대규모 또는 강력한 컴퓨터의 성능을 이용하여 효과적인 서비스를 가능하게 한다. 그러나 이런 경우에 주로 문제시 되는 것은 전송되는 음성신호가 이동통신채널에 의해 쉽게 왜곡될 수 있다는 것이다.

책임저자: 장 준 혁 (jhchang@ece.ucsb.edu)
Electrical and Computer Eng. Univ. of California, Santa Barbara, CA, USA, 93117
(전화: 805-685-0753; 팩스: 805-893-3262)

실제로 이동통신환경에서 음성신호가 왜곡될 때 음성인식의 성능이 급격히 저하되는 문제를 해결하기 위한 광범위한 연구가 진행되어져 왔다. 가능한 방법은 기존의 음성인식거부방법을 사용하는 것인데 관련 연구방법들은 특징벡터 추출, 인식과정 등을 거치기 때문에 많은 계산량을 요구하게 되므로 응답시간이 늦다[1,2]. 게다가 기존의 연구들 대부분이 현재의 이동통신채널에 대한 고려가 부족하기 때문에 인식성능향상이 제한적일 수 밖에 없다. 제시한 이유로 본 연구에서는 이동통신환경하에서 적은 계산량으로 채널에 의해 심각히 오염된 음성신호를 거부하는 알고리즘을 제안한다. CDMA 이동통신 환경에서의 많은 음성데이터베이스를 바탕으로 음성왜곡은 크게 세 가지의 부류로 일어난다는 것을 발견하고 본 논문에서는 전형적인 CDMA채널 왜곡된 음성을 거부하기 위한 전처리-거부방법을 제안한다. 음성인식실험

을 통해서 제안된 방법이 적은 계산량을 가지고 효과적으로 왜곡된 음성신호를 전처리 거부할 수 있다는 것을 알 수 있었다.

II. 이동통신환경에서의 음성신호

이동통신환경에서 음성신호는 신호원만이 아니라 통신채널에 의해서 심각하게 왜곡되어질 수 있다[3,4]. 먼저 이동통신채널의 환경에 대한 이해를 위해 전송채널의 특성에 대해 고찰한다. 전송 중에 채널과 간섭신호에 의해 에러가 빈번히 발생하므로 음성신호처리입장에서 보다 심도있는 전송상의 경로에 대한 이해가 필요하다 [5]. 실제 대도시환경에서 많은 전송상의 경로가 존재하게 되며 이것이 반사, 흡수, 산란등을 유발하게 된다. 특히, 제시된 문제는 경로상에서 송신단과 수신단이 가시경로(line of sight, LOS)상에 있지 않을 때 발생하게 된다. 이전 조사에 따르면 송신신호에서의 왜곡을 일으키는 주된 이유는 크게 두 가지로 설명된다. 첫째, shadow fading은 수신단이 이동중에 있을때 흔히 일어나는데 수신자가 서로 높이가 다른 빌딩이나 공터, 교차로 등을 지나갈때 전송되는 신호의 크기가 위치에 따라 급격이 변하게 되는 현상이다. 둘째, 전송신호의 크기가 좁은 위치의 구역에서 짧은 시간을 단위로 급격히 변화하며 이 현상을 fast fading이라고 한다.

제시된 통신채널상의 효과를 분석하기 위해서 자동응답시스템 (interactive voice response, IVR) 이 실험에 채택되었다. 한국에서 상용화된 CDMA 이동통신채널

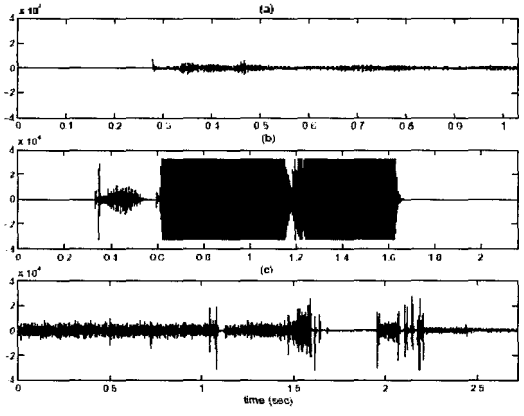

그림 1. CDMA환경에서 채널 왜곡된 음성신호의 전형적인 예 (a) TYPE 1 (b) TYPE 2 (c) TYPE 3

Fig. 1 Typical example for channel distorted speech signal in CDMA environements (a) TYPE 1 (b) TYPE 2 (c) TYPE 3

(code division multiple access)을 통해 전송되어 진 음성신호를 자동응답시스템을 통해 저장하여 데이터베이스화 하였다. 음성데이터는 건물안, 밖, 거리, 정지 및 이동중인 차와 같은 다양한 환경으로 채집되었다. 실제 채집된 음성데이터가 음성인식오류를 일으키는 데는 다음과 같은 대표적인 세 가지의 유형으로 정리할 수 있었다.

TYPE 1 : 알아들을 수 없을 정도의 파형감쇄
TYPE 2 : 파형 클리핑
TYPE 3 : 파형 손실

전형적인 세 가지의 예에 대한 실제 예가 그림 1에 나타나있다. 결론적으로, fading과 같은 이동통신환경에서의 채널왜곡 요인이 음성신호의 왜곡을 초래하고 이것은 자동음성인식의 성능을 급격히 저하시킨다는 것이다. 실제로 위에 분류된 음성신호들은 자동음성인식기에 입력되었을때 거의 잘못 인식되거나 기존의 거부모듈에 의해 거부되었다. 그러나 기존의 거부방법들은 특징벡터추출, 확률계산, 검색등의 고도의 계산량을 필요로 하므로 많은 인식채널을 사용하는 서버중심의 자동인식시스템에서는 적합지 않다. 본 연구에서는 언급된 계산상의 로드를 줄이고 사용자에게 즉각적인 인식오류를 전달할 수 있는 음성신호의 prosody정보만을 이용한 전처리-거부방법을 제안한다.

III. 음성인식을 위한 전처리거부방법

음성신호는 성도 (vocal cord)로부터 여기되어서 나오는 생성과정에서 알 수 있듯이 주기적인 구조를 가지고 있다. 따라서, 음성신호의 주기성은 음성을 분별하기 위한 주된 특성중의 하나이다. 여기서 음성신호여부를 판단하기 위해 정규화된 피치 상관도 (normalized pitch correlation) 과 피치를 사용한다[6]. 즉, 여기서 사용되는 알고리즘의 주요한 동기는 음성신호가 이동통신채널에 의해 심각히 왜곡되는 경우, 음성신호의 주기성이 쉽게 깨어진다는 사실에 기반한다. 실제 구현을 위해 10 ms의 프레임의 음성신호를 처리하였고 샘플링 주파수는 8000 Hz이었다. 정규화된 피치상관도는 선형예측방법 (linear prediction, LP)에 의해 얻어지며 다음과 같이 구해진다[7].

$$R_m(\tau)=\frac{\sum_{n=0}^{N-1}s(n)s(n+\tau)}{\sqrt{\sum_{n=0}^{N-1}s^2(n+\tau)}} \quad (1)$$

여기서 m 은 프레임 (10 ms) 첨자이고, N은 윈도우사이즈, τ은 시간쉬프트를 각각 나타낸다. 식 (1)을 이용하여, 개구간 피치지연을 계산하면

$$\hat{p}=\arg\max_{\tau} R_m(\tau). \quad (2)$$

그림 2에서 채널 왜곡된 대표적인 음성신호 세가지에 대한 개구간 피치를 표시하고 있다. 정규화된 피치상관도는 다음과 같이 정의된다.

$$R_p(m)=R_m(\hat{p}) \quad (3)$$

여기서 $\hat{p}$는 (2)에서 결정된 개구간 피치이다. 채널 왜곡된 대표적인 음성신호 세 가지에 대한 정규화된 피치상관도를 그림 3에서 표시하였다.

연속으로 이웃하는 음성프레임 피치파라미터를 조사하기 위해 이전의 5개의 정규화된 피치상관을 R_p와 P를 계산하는데, 실제로 $R_p=[R_p(m-4),R_p(m-3),\cdots,R_p(m)]$와 $P=[p(m-4),p(m-3),\cdots,p(m)]$이고 여기서 $R_p(m)$과 $p(m)$은 각각 정규화된 프레임 m에서의 피치상관도과 피치를 나타낸다. 우리의 접근방법은 위에서 정의되는 두 파라미터를 근간으로 주기성을 조사하는 방법을 이용한다. 정규화된 피치상관의 running평균을 다음과 같이 구한다.

$$\overline{R}_p(m)=\lambda_R\overline{R}_p(m-1)+(1-\lambda_R)\mu_{R_p} \quad (4)$$

여기서 μ_{R_p}는 R_p의 정규화된 피치상관의 평균값이고, λ_R(=0.8)은 long-term 스무딩 파라미터이다. 각 프레임에서의 결정룰은 다음과 같이 이루어진다.

if $\overline{R}_p(m)>\tau_{\mu_1}$ then FLAG = 1

else if $\{\overline{R}_p(m)>\tau_{\mu_2}$ and $\sigma_P<\tau_\sigma\}$

then FLAG = 1

else FLAG = 0

여기서, σ_p는 P에서 계산된 피치값들의 표준편차이고 τ_{μ_1}(=0.63), τ_{μ_2}(=0.45), τ_σ(=1.30)은 실험적으로 최적화된 문턱값이다. 이 파라미터들의 목적은 왜곡되지 않은 음성신호의 준상관적 (quasi-stationary)인 특성을 고려하기 위한 것이다.

결정된 FLAG값에 근거하여 피치 연속성을 다음과 같이 설정한다.

if {FLAG = 1 and VAD = 1}

then $C_{pc}=C_{pc}+1$

else $C_{pc}=C_{pc}$

여기서 VAD는 음성이 존재하느냐 (=1) 존재하지 않느냐 (=0)에 대한 프레임상태에 대한 이진 결정값이며, 실제로 ASR의 과정의 끝점검출기로부터 전송되어 진다. 실제로 본 실험에서는 3GPP2의 SMV (selectable mode vocoder)의 음성검출기루틴을 채용하였다[6]. 채널왜곡

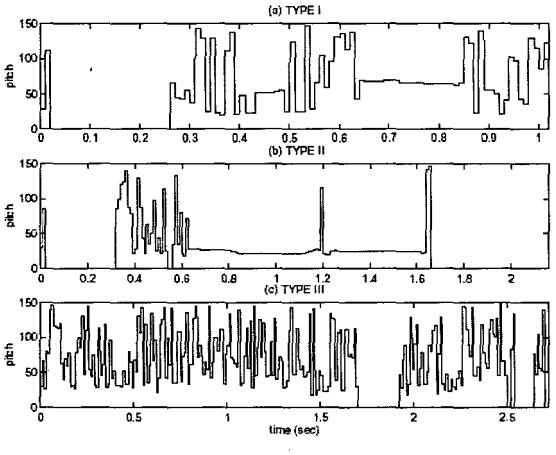

그림 2. 그림 1의 채널 왜곡된 음성신호에 대한 피치 괘적
Fig.2 Pitch contour for channel distorted speech signal in Fig. 1.

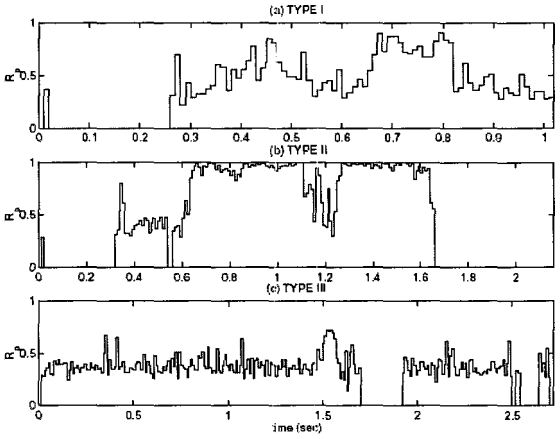

그림 3. 그림 1의 채널 왜곡된 음성신호에 대한 정규화된 피치상관도의 예
Fig. 3 Normalized pitch correlation for channel-distorted speech signal in Fig. 1.

된 음성신호 (TYPE1, 2, 3)와 채널왜곡되지 않은 음성신호를 나타내는 주어진 두 가설 H_0와 H_1로부터 전처리-거부방법의 최종 결정룰은 다음과 같이 구해진다.

$$C_{pc} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{5}$$

여기서 η는 결정룰에 대한 문턱값이다. 실제 상황에서는 검출성공과 검출실패를 고려해 신중히 결정하여야 한다. 여기서 중요한 것은 식(5)에서 C_{pc}는 음성검출기에 의해 충분히 긴 묵음이 검출될 경우마다 0으로 리셋하여야 하며 실제로는 0.15초로 세팅되어졌다.

IV. 실험 및 결과고찰

제안된 전처리-거부방법을 음성데이터베이스에 대해 평가하였다. 먼저, CDMA이동통신채널을 통해 전송된 음성신호를 저장하여 기준결정을 내렸다. 즉, 건물안, 밖, 차 등의 환경에서부터의 음성신호를 통신채널을 통해 저장토록 했다. 이 송신단의 사람은 사람의 이름을 즉시 발성하게 했으며 전체 1100개의 이름목록이 사용되어져서 2345개의 파일이 기록되었다. 먼저, 채널왜곡에 따른 음성인식의 실패유무를 알아보기 위해 각 파일은 인식기에 적용되어 진 후, 인식성공한 경우와 실패한 경우의 두 부류로 나누었다. 인식실패한 파일들은 채널에 의해 왜곡된 경우와 그렇지 않은 경우로 수동으로 레이블링 되었다. 결국 2126파일은 채널에 의해 오염되지 않은 부류로 219파일은 오염된 신호로 분류되어졌다. 구체적으로 105개의 파일이 TYPE1, 60개의 파일이 TYPE2, 54개의 파일이 TYPE3에 할당되었다. 실제로 인식에 성공한 경우에서는 TYPE1과 TYPE3에 해당하는 음성파형의 경우가 없었으나 TYPE2의 경우에는 음성인식이 성공된 경우도 존재했다. 실제로 음성인식에 성공과 실패한 모든 경우에 TYPE2를 조사한 결과 전체 TYPE2에 해당하는 음성신호에서 약 60%가 음성인식에 실패한 것으로 판명되어 실제 TYPE1과 TYPE3에 해당하는 채널상태가 음성인식 시스템에 가장 열악한 환경임을 알 수 있었다.

자동음성인식 시스템은 hidden Markov models (HMMs)[8]을 기반으로 하여 서울대학교에서 자체 개발된 것으로, 35명의 화자가 학습데이터를 구성하기 위해서 CDMA상의 오염되지 않은 상태에서 녹음되어진 38,500의 이름단어가 사용되어 졌으며 또 다른 별도의 10명의 화자로부터 녹음받은 데이터를 바탕으로 실험한 음성인식결과는 94.5%였다.

녹음된 입력음성은 8 kHz샘플링되어 30 ms로 세크먼트되었다. 각 이름은 triphone으로 모델링되어 각 폰은 three-state left-to-right HMM without skips과 two mixture components으로 나타내어 졌다. 묵음모델은 각 이름의 양쪽에 위치시켜서 유연한 워드경계를 이루었다. 실험에서 사용한 특징벡터는 8차 Mel 주파수 cepstrum계수와 log energy, 그것의 일차편차를 사용했다.

위의 인식과정을 거쳐서 잘못 인식된 음성신호는 다시 채널 왜곡되거나 왜곡되지 않은 음성신호로 수동 구별되었다. 수동 구별된 기초 자료를 왜곡여부의 근거자료로 삼아 전체 파일을 전처리 거부 시스템에 인가하여 제안된 결정룰이 왜곡된 음성신호를 음성인식에 실패한 음성신호인지 여부에 대하여 판단하게 하였다. 즉, 이것은 전체 음성신호파일을 인식기에 인가하지 않고 단지 새로이 제안된 전처리거부 알고리즘만으로 채널왜곡에 따른 인식실패를 미리 예측, 판단한다는 점이 다르다.

여기서 왜곡된 음성신호로 제대로 판단한 확률을 P_d, 그렇지 않은 확률을 P_f로 나타내기로 하자. 그림 4에서 문턱값 η 에 따른 P_d과 P_f사이의 trade-off를 나타내주는 수신 동작 특성 (Receiver operating characteristic, ROC)을 보여주고 있다.

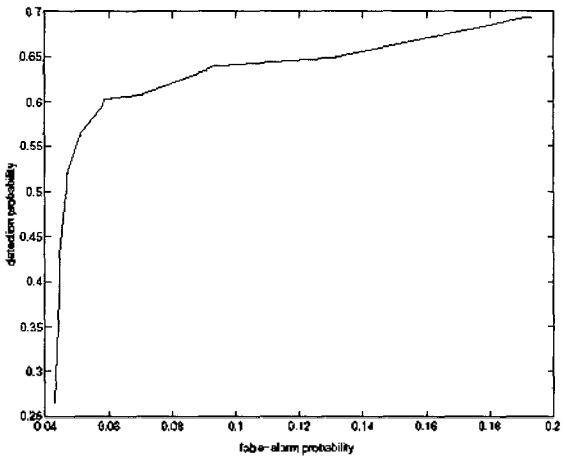

그림 4. 제안된 전처리거부방법에 대한 수신동작특성 (Receiver operating characteristics)

Fig. 4 ROC (receiver operating characteristics) for proposed pre-rejection algorithm.

그림 4에서 보듯이 제안된 알고리즘은 채널에 의해 왜곡된 음성신호에 대해 P_f가 5%로 고정된 경우에 대해 P_d>60% 정도의 성능을 보였다.

V. 결 론

본 논문에서는 음성인식기의 전처리기법으로서 CDMA 이동통신채널에 의해 오염된 음성신호을 분석하여 세 가지의 대표적인 채널 왜곡된 음성신호로 분류하였다. 분석된 채널왜곡된 음성신호를 바탕으로 하여 간편하게 전처리-거부할 수 있는 방법을 제안하였는데, 실제로 이것은 왜곡되지 않은 음성신호의 준주기성에 기반한 것으로 적은 계산량으로 구현되어질수 있고, 쉽게 상용시스템에 적용할 수 있는 장점이 있다.

참 고 문 헌

1. M. Weintraub, "LVCSR log-likelihood ratio scoring for keyword spotting," Proc. of IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.297-300, 1995.
2. R. C. Rose, "Discriminant wordspotting techniques for rejecting non-vocabulary utterances in unconstrained speech", Proc. of IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 105-108, 1992.
3. J. -H. Chang and N. S. Kim, "Speech enhancement : new approaches to soft decision," IEICE Trans. Inf. and Syst., 27, E84-D, pp. 1231-1240, Sep. 2001.
4. Y. Okumura, E. Ohmori, T. Kawano and K. Fukuda, "Field strength and its variability in VHF and UHF land-mobile radio service," Review of the Electrical Commnuication Laboratory, Vol. 16(9-10), Sep.-Oct. 1968.
5. H. L. Bertoni, Radio propagation for modern wireless systems (Prentice Hall, 2000).
6. TIA/IS-893, Version 4.3 of the SMV *algorithm description* text (3GPP2, 2001).
7. A. M. Kondoz, *Digital speech : coding for low bit-rate communications systems* (John Wiley & Sons, 1994).
8. L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition* (Prentice Hall, Englewood Cliffs, NJ, 1993).

저자 이력

• 장 준 혁 (Joon-Hyuk Chang)

1973년 9월 16일생
1998년 2월: 경북대학교 전자공학과 학사
2000년 2월: 서울대학교 전기공학부 석사
2004년 2월: 서울대학교 전기컴퓨터 공학부 박사
2000년 3월~현재: 쥬넷더스 연구소장
2004년 5월~현재: 캘리포니아 주립대학, 산타바바라 박사후 연구원

• 김 남 수 (Nam Soo Kim)

1965년 10월 18일생
1988년 2월: 서울대학교 전자공학과 학사
1990년 2월: 한국과학기술원 전기 및 전자공학과 석사
1994년 8월: 한국과학기술원 전기 및 전자공학과 박사
1993년 5월~1998년 2월: 삼성종합기술원 전문연구원
1998년 3월~현재: 서울대학교 전기컴퓨터공학부 부교수