LSP 파라미터를 이용한 발성측정법

장 경 아, 배 명 진
숭실대학교 정보통신공학과
전화 : (02) 824-0906 / 팩스: (02) 820-0018

On a Study of Measurement Method of Utterance Velocity for the Reduction of Transmission Rate in CELP Vocoder.

KyungA JANG, MyungJin BAE
Dept. Information and Telecommunication Engr., Soongsil University
E-mail : kajang@assp.ssu.ac.kr

Abstract

Speaking Rate has variety depends on the situation and habit of speakers. It has been many studied about speaking rate in speaker recognition. The study of speaking rate in speech recognition is one of considerable matter when it is recognized the speakers and it is measured by many speech data base and complicate estimation for accuracy. In this paper, conventional vocoder process the speech signal when encoding and transmitting without regard to speaking rate so in order to apply the speaking rate for vocoder it should be considered the simpler algorithm and less computation amount than the conventional method of speaking rate used in speech recognition. We proposed the speaking rate algorithm which is used the simple parameter with Line Spectrum Pair (LSP). The proposed peaking rate method is measured by the information of LSP in speech. We measured the variety rate of phenomenon about utterances which have different velocity, respectively. As a result, It has distinct variation rate of phenomenon between utterances uttered fast and slow and the rate is 42.8% higher in case of uttered fast than in case of uttered slow.

I. 서 론

기존의 보코더에서 입력된 음성신호를 부호화하여 전송시 음성 발성 속도에 대해서 고려하지 않는다는 점을 감안하여 좀 더 전송율을 낮추기 위한 파라미터로써 발성속도를 측정하고자 한다. 일반적으로 발성속도 측정법은 음성 합성시 이나 음성인식에서 사용하고 있는데, 이 방식은 낮은 전송율로 전송이 가능해야 하고, 간단한 알고리즘으로 복잡도가 낮어야 하는 부호화기에서는 적합하지 않는 방법으로 간단한 파라미터를 사용하면서, 복잡하지 않은 발성속도 측정법을 제안하고자 한다.

II.1 기존의 발성속도 측정

음소 구간의 변화의 중요한 요소는 발성속도의 변화이다. 발성 속도는 포괄적인 측정으로 단위 시간당 음소의 평균 개수로 정의된다. ARPA WSJ 데이터 베이스에서 수집된 문장들은 평균 120 음소들의 길이로 이루어져 있어, 전체 문장에서 측정된 발성속도는 발성에 따라 변화는 무시된다. 따라서 음소에서의 측정된 음소 세그멘트의 구간을 바탕으로 하는 각각의 음소 세그멘트에 대해 발성속도를 측정한다. 발성속도는 다음과 같이 정의된다.

$$r_i = \frac{\sum_{k=-M1}^{M} d_{i+k}}{\sum_{k=-M}^{M} ttl_{i+k}} \qquad (2.1)$$

여기서 r_i 은 i 번째 음성 세그멘트의 발성 속도이고, d_{i+k}와 ttl_{i+k}은 각각 (i+k)의 음성 세그멘트의 구간 확률 분포에서 각각 관찰된 구간과 예정된 구간들이다. 이 정의는 하나의 음소 세그멘트에서 전체 문장까지의 어떤 길이에서든 적절한 변수 M 에 의해

서 발성속도 계산 윈도우를 융통성 있게 조절할 수 있도록 한다. 음성 세그멘트 각각에 대해 발성속도 측정은 두 개의 다른 구간 모델들을 사용하므로써 발성속도 정보에 대해 잇점을 얻는다. 훈련 데이터를 다른 셋(set)에 묶어놓기 위해 발성속도 정보를 사용하고 각 셋(set)안에서 구간 모델들에 대해 히스토그램과 확률밀도 함수를 계산한다. 음소 세그먼트들의 구간은 발성속도의 함수라 가정하고, 하나의 평준화된 구간 모델을 만들기 위해 이 가정을 사용한다. 두 개의 접근 방식은 어떠한 발성속도 정보를 고려없이 훈련 데이터에서 일어날 수 있는 모든 구간 모델들에 대해 히스토그램을 생성하기 위해 필요하다. 이 통계는 음성 세그먼트를 계산하기 위한 식 (2.1)에서 쓰여진 모델의 평균 구간 값의 계산에서 매우 필요하다[2].

III. LSP 파라미터 추출

LSP 파라미터를 추출하기 위해서 먼저 LPC(Linear Predictive Coding)분석이 이루어져야 한다[3].

$$H(z)=1/A_p(z) \tag{3.1}$$

$$where \quad A_p(z)=1+\sum_{k=1}^{p}a_k z^{-k} \tag{3.2}$$

$H(z)$는 LPC 필터이고 p는 필터의 차수이다.
LSP 파라미터를 유도하기 위해서 PARCOR(Partial Correlation) 필터를 이용해서 식(3.1)과 식(3.2)를 표현하면 다음과 같다.

$$\begin{aligned} A_{p-1}(z)&=A_p(z)+k_pB_{p-1}(z)\\ B_p(z)&=z^{-1}[B_{p-1}(z)-k_pA_{p-1}(z)] \end{aligned} \tag{3.3}$$

여기서 $A_0(z)=1$고 $B_0(z)=z^{-1}$ 이고

$$B_p(z)=z^{-(p+1)}A_p(z^{-1}) \tag{3.4}$$

PARCOR 구조는 손실이 없는 음파관에서 음파의 전달로 이해된다. 시스템은 단지 역방향(backward) 에너지 모양에서 Z 종점에서 손실이 있다. 이러한 음관은 Z 종점의 출력이 $k_{p+1}=\pm1$의 경로를 통해 입력의 종점으로 귀환될 때 완전한 무손실이 된다. 각각의 공명 값인 Q는 무한해지고 에너지 분포 스펙트럼은 몇 개의 선 스펙트럼에 집중된다[3][4]. $k_{p+1}=-1$ 조건의 귀환은 입력종점에서 완전히 폐쇄되고 $k_{p+1}=+1$ 은 무한 자유공간상으로 개방된다[1][6]. $k_{p+1}=\pm1$ 인 전달함수를 $P_{p+1}(z)$와 $Q_{p+1}(z)$ 로 나타내면:

$$\begin{aligned} &k_{p+1}=1 \text{ 일때}, \quad P_{p+1}(z)=A_p(z)-B_p(z)\\ &k_{p+1}=-1 \text{ 일때}, \quad P_{p+1}(z)=A_p(z)+B_p(z) \end{aligned} \tag{3.5}$$

$$\Rightarrow A_p(z)=\frac{1}{2}[P_{p+1}(z)+Q_{p+1}(z)] \tag{3.6}$$

두 개의 근($k_{p+1}=\pm1$)을 알고 있으므로 $P_{p+1}(z)$의 $Q_{p+1}(z)$의 차수를 줄일 수 있다. 즉,

$$\begin{aligned} P'(z)&=\frac{P_{p+1}(z)}{(1-z)}\\ &=A_0z^p+A_1z^{(p-1)}+\cdots+A_p \end{aligned} \tag{3.7}$$

$$\begin{aligned} Q'(z)&=\frac{Q_{p+1}(z)}{(1-z)}\\ &=B_0z^p+B_1z^{(p-1)}+\cdots+B_p \end{aligned} \tag{3.8}$$

조건 : $A_0=1, \quad B_0=1$ (3.9)

$$A_k=(a_k-a_{p+1-k})+A_{k-1} \tag{3.10}$$

$$B_k=(a_k-a_{p+1-k})-A_{k-1} \quad \text{for} \quad =1,\cdots p$$

LSP는 $0\le\omega_i\le\pi$인 범위에서 $P'(z)$와 $Q'(z)$ 통해 얻어진 근의 각(angular) 위치를 나타낸다. LSP는 다음과 같은 두가지 성질을 가지고 있다.

첫째, $P'(z)$와 $Q'(z)$는 단위원 상에 놓여 있다.

둘째, $P'(z)$와 $Q'(z)$의 근들이 단위원 상에 번갈아 나타난다.

Real root method

일반적으로 $P'(z)$와 $Q'(z)$의 다차 방정식의 해를 구하는 방식에 따라 여러 가지 변환법이 개발되었다. 그러나 이러한 변환법 중 real root 방법이 비교적 간단하고 이해하기 쉬워 주로 사용되어지고 있다[3].

$P'(z)$와 $Q'(z)$의 계수는 대칭적이기 때문에 식(3.7)의 차수는 $p/2$로 줄어든다.

$$\begin{aligned} P'(z)&=A_0z^p+A_1^{p-1}+\cdots+A_1z^1+A_0\\ &=z^{p/2}[A_0(z^{p/2}+z^{-p/2})+\\ &\quad A_1(z^{(p/2-1)}+z^{-(p/2-1)})+\cdots+A_{p/2}] \end{aligned} \tag{3.11}$$

$$\begin{aligned} Q'(z)&=B_0z^p+B_1^{p-1}+\cdots+B_1z^1+B_0\\ &=z^{p/2}[B_0(z^{p/2}+z^{-p/2})+\\ &\quad B_1(z^{(p/2-1)}+z^{-(p/2-1)})+\cdots+B_{p/2}] \end{aligned} \tag{3.12}$$

모든 근이 단위원 상에 있기 때문에, 단지 아래와 같이 정의하고 단위원 상에서 식(3.11)의 값을 구할 수 있다.

$$Let \quad z=e^{j\omega} \quad then \quad z^1+z^{-1}=2\cos(\omega) \tag{3.13}$$

$$\begin{aligned} P'(z)=2e^{jp\omega/2}[&A_0\cos(\frac{p}{2}\omega)+\\ &A_1\cos(\frac{p-2}{2}\omega)+\cdots+\frac{1}{2}A_{p/2}] \end{aligned} \tag{3.14}$$

$$Q'(z)=2e^{jp\omega/2}[B_0\cos(\frac{p}{2}\omega)+ B_1\cos(\frac{p-2}{2}\omega)+\cdots+\frac{1}{2}B_{p/2}] \quad (3.15)$$

$x=\cos\omega$ 를 대입해서 식(2.14)와 식(2.15)을 x에 대해서 풀 수 있다. 예를 들어서 $p=10$ 이면 다음과 같이 얻어진다.

$$P'_{10}(x)=16A_0x^5+8A_1x^4+(4A_2-20A_0)x^3+(2A_3-8A_1)x^2+(5A_0-3A_2+A_4)x+(A_1-A_3+0.5A_5) \quad (3.16)$$

유사하게,

$$Q'_{10}(x)=16B_0x^5+8B_1x^4+(4B_2-20B_0)x^3+(2B_3-8B_1)x^2+(5B_0-3B_2+B_4)x+(B_1-B_3+0.5B_5) \quad (3.17)$$

LSP는 식(3.18)에 의해서 구해진다.

$$LSP(i)=\frac{\cos^{-1}(x_i)}{2\pi T},\quad \text{for}\quad 1\le i\le p \quad (3.18)$$

이 방법은 다른 변환 방법보다 비교적 간단하나 계산시간이 어느 정도 걸릴지 예상할 수 없다는 단점이 있다[6].

IV.발성속도에 따른 음소변화율 측정

본 논문에서 제안하고자 하는 알고리즘은 CELP형 부호화기에서 음성신호를 고려하지 않는다는 점을 감안하여, 저전송율 부호화기에 맞도록 복잡하지 않는 알고리즘으로 발성속도율 구하는 방법을 제안하고자 한다. 여기서 사용하고자 하는 파라미터는 에너지와 LSP 계수이다. 먼저 에너지 임계값 및 묵음 구간의 LSP 계수 설정을 설정한다.

$$Ene_i=\frac{\sum_{n=0}^{N-1}s_{i,t}^2[n]}{N},\quad i=0,1,2 \quad (4.1)$$

$$NLSP_k=\sum_{n=0}^{2}LSPvect_{i,t},\quad i=0,1,2,\ldots,10 \quad (4.2)$$

여기서 N은 240이며, $s_i[n]$은 현재 프레임의 t의 입력신호이며 LSP는 현재 프레임에서 구한 LSP 계수들이다. 위의 파라미터를 이용하여 다음과 같은 에너지 임계값과 묵음의 평균 LSP 계수들을 계산한다.

$$EneThr=mean(Ene) \quad (4.3)$$

$$LSPave_k=NSLP_k,\quad k=1,2,..,10 \quad (4.4)$$

에너지 임계값을 넘는 경우와 에너지 임계값을 넘는 경우지만 입력 경우가 낮은 SNR을 갖는 경우를 고려하기 위해 LSP 파라미터를 이용하여 판정을 수행한다. 묵음 구간의 LSP 계수들 사이에는 일반적으로 등간격을 가지고 있지만 음성이 존재하는 경우는 포만트가 위치하는 주파수영역에 LSP들이 많이 존재하는 특징이 있다 [9]. 즉 묵음구간에서 구한 LSP 계수들과 음성이 존재하는 LSP 계수들 사이의 오차를 구하면 그 값이 크게 되지만 묵음 구간의 LSP 계수들 사이의 오차는 상당히 적게 된다. 따라서 LSP 계수들 사이의 오차를 이용하면 음성의 존재유무를 판정할 수 있게 된다. LSP 계수들 사이의 거리는 다음과 같이 구할 수 있다[2].

$$LSPdist=\sqrt{\sum_{i=1}^{10}\{LSP_t(i)-LSPave(i)\}^2} \quad (4.5)$$

위에서 구한 에너지와 LSP 거리값이 미리 설정된 각각의 임계값보다 작은 경우 묵음 구간으로, 그렇지 않은 음성이 있는 구간으로 설정하게 된다. 음성이 있는 구간으로 판정된 프레임에서는 각 10차의 LSP 계수를 구하게 된다. 구해진 LSP 계수는 인접 프레임간의 변화도를 보기 위해 다음 프레임의 LSP 계수값에서 현재 프레임의 LSP 계수값의 차를 구한다.

$$LSPdiff(n)=\sum_{i=0}^{p}|LSP_{n+1}(i)-LSP_n(i)|,\ i=1,2,\ldots,P \quad (4.6)$$

$$LSPdiff_{mean}(i)=\frac{1}{3}\sum_{i=1}^{n+2}LSPdiff(n),\ n=nthFRAME \quad (4.7)$$

한정된 인접 프레임간의 변화도의 평균값 $LSPdiff_{mean}$ 을 가지고, 1초당 처리되는 프레임 수를 구하여 현재 프레임이 $FrPersec$ 을 넘는 경우와 그렇지 않은 경우를 달리하여 처리한다. 현재 프레임이 $FrPersec$을 넘지 않은 경우이면서 LSP 변화도 문턱값을 넘는 경우에는 다음과 같이 측정한다.

$$SpRate(n)=\sum_{n-FrPerSec}^{n}\text{if}\,(LSPdiff_{mean(n)}>thDiff) \quad (4.8)$$
$$*of(\ )\rightarrow(\ )\text{을 만족하면 1, 그렇지않으면 0}$$

1이라 만족된 경우에 발성속도율을 구한다. 구하는 식은 다음과 같다.

$$SpRate(n)'=FrPerSec/n\times SpRate(n),\ n=nth\ frame \quad (4.9)$$

현재 프레임이 $FrPersec$을 넘는 경우이면서 LSP 변화도 문턱값을 넘는 경우이면 다음과 같이 구한다.

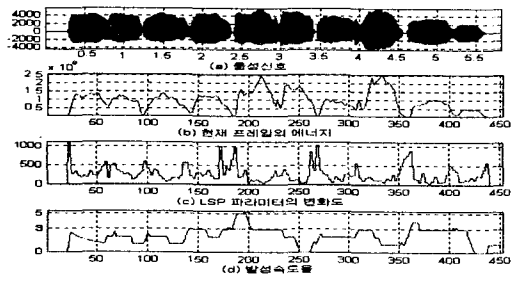
$$SpRate(n)'=\frac{diff(n-FrPerSec)}{diff_{mean}},\ n=nth\ frame \quad (4.10)$$

V. 실험 및 결과

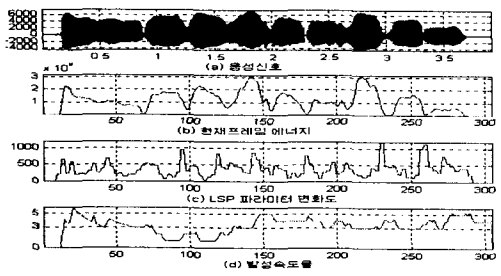
제안한 방법을 실험하기 위해서 먼저 IBM PC(233 MHz)에 마이크 입력이 가능한 A/D 변환기를 인터페이스 하였다. 음성시료는 남자와 여자가 연구실 환경(30dB의 SNR)에서 발성한 음성을 8kHz로 표본화하고 16bit로 양자화하여 사용하였다. 발성한 문장은 다음과 같다. 음성시료는 발성속도를 각각 다르게 하여 같은 문장을 발성하였다.

발성1) "아야어여우유으이"
발성2) "여기는 음성통신 연구실입니다"
발성3) "일이삼사오육칠팔구십"
발성4) "아름다운 가을입니다"

제안한 음소변화율 알고리즘은 C언어로 구현하여 각각 걸리는 시간과 각 발성문장에 대한 음소변화율을 측정하였다. 그림은 음소변화율을 구하기 위해 각 파라미터들을 구하는 과정이다. (c)는 인접 프레임간 각 차수의 LSP 파라미터의 변화도를 나타낸 것이다. 음소가 지속되는 경우에는 LSP 변화도가 적은 반면에, 음소가 변화는 경우에는 LSP의 변화도의 값이 큰 것을 볼 수 있다. (d)는 음소 변화율을 나타낸 것이다. 이 음소변화율의 측정치에서 빠르게 발성한 경우가 느리게 발성한 경우보다 약 42.8%가 높게 나왔다. 이 결과는 발성속도에 따라 다른 변화율을 가진다는 것을 알 수 있고, 발성 속도에 따라 빠르게 발성한 경우에는 느린 발성보다 변화율이 높다는 것을 알 수 있었다.



4-1. 빠르게 발성한 경우의 음소변화율



4-2. 느리게 발성한 경우의 음소변화율

VI. 결 론

기존의 보코더에서 입력된 음성신호를 부호화하여 전송시 음성 발성 속도에 대해서 고려하지 않는다는 점을 감안하여 좀 더 전송율을 낮추기 위한 파라미터로써 발성속도를 측정하고자 한다. 일반적으로 발성속도 측정법은 음성 합성시 이나 음성인식에서 사용하고 있는데, 이 방식은 낮은 전송율로 전송이 가능해야 하고, 간단한 알고리즘으로 복잡도가 낮어야 하는 부호화기에서는 적합하지 않는 방법으로 간단한 파라미터를 사용하면서, 복잡하지 않은 발성속도 측정법을 제안하고자 한다. LSP 파라미터가 가지고 있는 정보를 이용하여 음성의 발성속도에 따른 음소변화율을 구한 결과 빠르게 발성한 경우가 느리게 발성한 경우보다 42.8% 높게 나왔다.

표 4-1. 시간당 음소변화율

	발성 시간		음성 발성속도 변화율	
	Fast	Slow	Fast	Slow
발성(1)	4.62	7.10	3.879	2.264
발성(2)	3.21	6.52	4.149	3.187
발성(3)	4.2	6.53	3.472	2.129
발성(4)	2.3	4.53	3.432	1.103
평균	3.58	6.10	3.733	2.170

참고 문헌

[1] 배명진, "디지탈 음성분석", pp.95-120, 홍영출판사, 1998. 4

[2] M.J.Russel, K.M.Ponting and M.J Tomlinson, Measure of local speaking-rate for automatic speech recognition, ELECTRONICS LETTERS, 13th May.1999

[3] A.N. Ince, Digital Speech Processing (speech coding, synthesis, and recognition), Kluwer Academic Publishers, 1992.

[4] John R. Deller, Jr., John G. Proakis, John H.L. Hansen, "Discrete-Time Processing of Speech Signals", pp.124-125, Maxwell Macmillan International, 1993.

[5] Sadaoki Furui, "Digital Speech Processing, Synthesis, and Recognition", pp129, MARCEL DEKKER, INC. 1991.

[6] A. M. Kondoz, "Digital Speech", pp. 84-92, John Wiley & Sons Ltd, 1994.

[7] B.S. Atal and J.R. Remde "A New Model of LPC Exitation for Producing Natural-Sounding Speech at Low Bit Rates", Proc.Int.Conf. on Acoust., Speech and Signal Processing, pp.614-617.,Apr.1982.