동적시간외곡에 기초한 본문의존형화자인식에서 한가지 류사도표준화방법

리은성, 리광일

경애하는 최고령도자 **김정은**동지께서는 다음과 같이 말씀하시였다.

《기초과학은 과학기술강국을 떠받드는 주추입니다. 기초과학이 든든해야 나라의 과학기술이 공고한 토대우에서 끊임없이 발전할수 있습니다.》(《조선로동당 제7차대회에서 한 중앙위원회사업총화보고》 단행본 40페지)

본문의존형화자인식은 발성자가 등록때 발성하는 본문과 인식때 발성하는 본문이 일치할것을 요구하는 화자인식이다. 본문의존형화자인식은 등록과 인식때에 발성내용이 일치할것을 요구하지 않는 본문독립형화자인식에 비해 등록과 인식때 발성음성의 길이가 짧고 정확성이 높은 우점을 가지고있는것으로 하여 많이 연구되고있다.

본문의존형화자인식에는 동적시간외곡(DTW), 가우스혼합모형(GMM), 벡토르량자화(VQ), 숨은 마르꼬브모형(HMM)에 기초한 방법들이 있다. 이 방법들은 다 등록때 발성자가 발성한 음성자료들을 리용하여 발성자의 등록모형을 구성한다. 등록모형을 구성하는데 리용할수 있는 훈련자료의 량(발성자가 같은 본문을 발성한 차수)이 작을 때에는 동적시간외곡에 기초한 방법의 성능이 가장 높다.[1]

등록때 발성자에게 같은 본문을 많은 차수 발성할것을 요구하는것은 사용자에게 불편을 주므로 우리는 동적시간외곡에 기초한 본문의존형화자인식방법에 대하여 연구한다.

선행연구[1, 2]에서는 본문의존형화자인식에서 멜주파수케프스트람결수(MFCC)특징이 여러가지 스펙트르특징중에서 가장 인식성능이 높다는것을 실험적으로 보여주었다.

인식은 음성으로부터 특징을 추출하고 등록된 모형들과 비교하여 류사도를 계산하고 계산된 류사도를 고정된 턱값과 비교하는 방법으로 진행된다. 그러나 발성자의 감정변화, 잡음, 감기 등 여러가지 요인의 영향으로 하여 음성에 대한 류사도의 변화가 대단히 크다. 이러한 류사도의 큰 변화는 고정된 턱값에 의하여 본인인가 아닌가를 판정할 때 인식성능을 크게 저하시키게 된다. 류사도의 변화를 줄이기 위한 방법으로서 특징수준에서의 처리, 모형수준에서의 처리 등이 연구되고있지만 그중에서 류사도표준화방법은 효과가 가장 높은것으로 인정되고있다. 대표적인 류사도표준화방법으로는 령표준화(Z-Norm)와 검사표준화(T-Norm)가 있다. 그러나 이러한 방법들은 대부분 본문독립형화자인식에 대하여 적용되여왔으며 본문의존형화자인식에 대하여서는 많이 연구되지 못하였다.

선행연구[3]에서는 가우스혼합모형과 숨은 마르꼬브모형에 기초한 본문의존형화자인식에서 류사도의 검사표준화에 대하여 연구하였다. 여기서는 화자인식체계를 리용하는 모든 사용자들이 같은 본문을 리용할것을 요구하므로 사용상 합리적이지 못한 결함이 있다.

선행연구[4]에서도 가우스혼합모형과 숨은 마르꼬브모형에 대한 류사도표준화에 대하여 연구하였으며 이 모형들은 원리적으로 등록을 위하여 많은 훈련자료들을 요구한다.

우리는 동적시간외곡에 기초한 본문의존형화자인식체계에서 류사도를 표준화하기 위한 한가지 방법을 제기하고 실험을 통하여 그 효과성을 검증하였다.

론문에서는 화자인식을 위한 특징으로서 선행연구들에서 그 효과성이 검증된 MFCC 특징을 리용한다.

먼저 발성자의 음성으로부터 MFCC벡토르렬의 계산과정에 대하여 보자.

발성자의 음성을 길이가 512인 프레임들로 나누고 매 프레임에 대하여 유무성판정을 진행한다. 프레임이 유성구간의 프레임인가 아닌가의 판정은 프레임의 에네르기와 령교차률을 리용하여 진행한다.

유성구간으로 판정된 매 프레임에 대하여 MFCC를 다음과 같이 계산한다.

프레임이 $\{x(n)\,|\,0\le n<512\}$와 같이 벡토르로 주어졌다고 하자.

전처리강조는 $y(n)=x(n)-0.95\cdot x(n-1),\ 0\le n<512$와 같은 일종의 고역강조려파처리이다. 여기서 $x(-1)$은 프레임시작전위치에서의 음성파형의 값이다.

우에서 얻어진 벡토르 $\{y(n)\,|\,0\le n<512\}$에 해밍창문을 곱한다.

$$y(n)=y(n)\cdot w(n),\ 0\le n<512$$

여기서 $\{w(n)\,|\,0\le n<512\}$는 해밍창문인데 다음과 같다.

$$w(n)=0.54-0.46\cdot\cos[2\pi n/(512-1)]\quad (n=0,\ \cdots,\ 512-1)$$

벡토르 $\{y(n)\,|\,0\le n<512\}$에 대하여 리산푸리에변환을 진행하여 $\{\hat{Y}(n)\,|\,0\le n<512\}$를

$$\hat{Y}(n)=\sum_{k=0}^{512-1}y(k)\cdot\exp\left(-2\pi i\frac{nk}{512}\right)\quad (n=0,\ \cdots,\ 512-1)$$

과 같이 계산한다.

MFCC를 계산하기 위하여 먼저 앞에서 구한 푸리에변환결수들의 절대값의 로그를 구하고 주파수축을 멜주파수척도로 변환한 다음 그 결과에 대하여 리산코시누스변환을 진행한다.

멜주파수척도는 공식

$$\mathrm{Mel}(f)=2\ 595\cdot\log(1+f/700)$$

에 의하여 정의된다.

우리는 리산코시누스변환결수의 앞 12개만을 모아 특징벡토르로 리용한다.

다음으로 동적시간외곡에 기초한 본문의존형화자인식체계에 대하여 고찰하자.

본문의존형화자인식체계는 등록과 인식으로 구성되여있다.

등록에서는 발성자의 음성으로부터 앞에서 서술한 특징추출을 진행하여 유성구간의 프레임렬로부터 MFCC벡토르렬을 계산하여 보관한다. 이것이 발성자의 등록모형이다.

화자인식을 위해서는 인식을 위하여 들어온 음성으로부터 등록에서와 같이 특징벡토르렬을 계산한 다음 이미 등록된 벡토르렬과 인식을 위한 특징벡토르렬을 동적시간외곡을 리용하여 대조를 진행한다.

등록된 벡토르렬과 인식용의 벡토르렬을 각각 $\{u_n\,|\,1\le n\le N_1\}$, $\{v_k\,|\,1\le k\le N_2\}$로 표시하고 이 두 렬에 대하여 동적시간외곡을 리용한 대조를 진행하기 위하여 다음과 같은 제한들을 주었다.

① (경사도조건) 경로의 경사도는 3을 넘을수 없고 1/3보다 작을수 없다.

② (경계조건) 경로의 시작점은 (1, 1), (1, 2), (2, 1)이, 끝점은 $(N_1,\ N_2)$, $(N_1-1,\ N_2)$, $(N_1,\ N_2-1)$이 될수 있다.

③ (국부조건) 마디점 $(i,\ j)$에 도달할수 있는 마디는 $(i-1,\ j)$, $(i,\ j-1)$, $(i-1,\ j-1)$이다.

론문에서는 두 특징벡토르사이의 다음과 같은 거리를 리용한다.

$$D(u,\ v)=1-\exp\left(-\frac{\|u-v\|^2}{\sigma^2}\right)$$

우의 제한조건과 거리를 리용하여 동적시간외곡에 의한 대조를 진행하여 최량경로 Tr, 경로의 길이 L, 두 벡토르렬사이의 거리 d를 계산한다.

Tr는 크기가 $2L$인 행렬이며 경로마디점의 i첨수와 j첨수를 보관한다.

이때 두 특징렬사이의 거리는 $\mathrm{sum}=\sum_{i=1}^{L}D(u_{Tr(i,\ 1)},\ v_{Tr(i,\ 2)})$, $d=\mathrm{sum}/L$과 같다.

다음 벡토르량자화모형을 리용한 류사도표준화에 대하여 론의하자.

본문의존형화자인식에서는 일반적으로 사용자마다 서로 다른 본문을 리용하므로 류사도의 표준화를 위하여 령표준화나 검사표준화를 리용하기가 어렵다.

론문에서는 사람의 음성특징의 일반적인 분포를 표현하는 벡토르량자화모형을 구성하고 이것을 리용하여 동적시간외곡에 기초한 류사도를 표준화하는 방법을 제기한다.

① 벡토르량자화모형의 구성

사람의 음성특징의 일반적인 분포를 표현하는 벡토르량자화모형을 구성하기 위하여 먼저 훈련모임을 구성한다. 남녀 각각 50명이 30s간 발성한 음성으로부터 앞에서와 같은 방법으로 MFCC벡토르들을 계산한다. 이때 매 사람이 발성하는 본문내용은 서로 달라도 된다. 계산된 특징벡토르들을 모두 모아 하나의 훈련모임을 구성한다.

다음 이 훈련모임에 K－평균무리짓기알고리듬을 적용하여 160개의 무리중심벡토르들을 구성한다.

이로부터 얻어진 무리중심벡토르들의 모임을 $\{C_i\,|1\le i\le 160\}$으로 표시하자.

매 중심벡토르는 12차원벡토르이다.

② 류사도표준화

먼저 인식을 위한 특징벡토르렬 $\{v_k\,|1\le k\le N_2\}$의 매 벡토르에 대하여 160개 무리중심벡토르와의 거리를 계산하고 그중에서 최소값을 구한다.

$$md(k)=\min_{1\le i\le 160}D(C_i,\ v_k),\ \ k=1,\cdots,\ N_2$$

다음 동적시간외곡에 기초한 대조에서 얻어진 최량경로를 리용하여 벡토르량자화모형에 대한 거리를 $vd=\sum_{n=1}^{L}\frac{md(Tr(n,\ 2))}{L}$와 같이 계산한다.

최종적으로 표준화된 류사도는 $\mathrm{sim}=d/\sqrt{vd}$에 의하여 계산한다.

이 류사도가 매 화자모형과 인식을 위한 발성음사이의 최종거리이다.

이 거리가 고정된 턱값보다 작으면 두 발성자는 같은 사람으로 판정하고 같거나 크면 서로 다른 사람으로 판정한다.

실 험 결 과

여기서는 론문에서 제기한 벡토르량자화모형을 리용한 류사도표준화방법의 효과성을 평가한다.

론문에서는 화자인식알고리듬들의 성능을 평가하기 위하여 5개의 단어 혹은 단어결합을 리용하여 5개 자료기지를 구성하였다. 매 자료기지는 24명의 음성자료를 포함하고 있다. 매 자료기지에는 매 사람당 서로 다른 시기에 5번 발성한 음성자료가 들어있다.

본문의존형화자인식의 성능평가절차는 다음과 같다.

매 자료기지의 첫시기에 발성한 음성을 리용하여 등록을 진행하고 나머지시기의 음성을 리용하여 대조를 진행한다.

론문에서는 2개의 대조방법에 대하여 성능평가를 진행한다.

방법 1은 본인인가 아닌가를 판정하기 위한 최종거리로서 론문에서 제기한 류사도표준화를 적용하지 않은 동적시간외곡에 기초한 대조에서 계산된 거리를 직접 리용하는 방법이고 방법 2는 론문에서 제기한 류사도표준화를 진행하여 얻어진 거리를 리용하여 판정을 진행하는 방법이다.

결국 두 방법은 류사도표준화를 진행하는가 안하는가에서만 차이가 있다.

매 자료기지에서 성능은 등오유률로써 평가하였다.

5개의 자료기지들에서 얻어진 성능은 표와 같다.

표. 5개 자료기지에서의 성능

자료기지 번호	방법	
	방법 1	방법 2
1	0.84%	0.48%
2	1.23%	1.02%
3	1.65%	1.15%
4	0.43%	0.24%
5	0.27%	0.21%

표에서 보는바와 같이 류사도표준화에 의하여 동적시간외곡에 기초한 본문의존형화자인식알고리듬의 성능이 크게 개선된다.

참 고 문 헌

[1] A. Ouzounov; Cybernetics and Information Technologies, 10, 1, 3, 2010.

[2] A. Bhise et al.; International Journal of Computer Science and Network, 2, 1, 6, 2013.

[3] M. L. Arslan et al.; Turk J. Elec. Eng. & Comp. Sci., 20, 2, 1277, 2012.

[4] M. Hebert et al.; Speech and Signal Processing, 1, 729, 2005.

주체107(2018)년 3월 10일 원고접수

A Similarity Normalization Method in the DTW-Based Text Dependent Speaker Recognition

Ri Un Song, Ri Kwang Il

We study the DTW-based text dependent speaker recognition. In order to reduce the similarity variation caused by various noises, we construct the universal background model using the VQ model and propose the method which normalizes the similarity using this model.

Key word: DTW