

# 语音反演远端监督学习模型研究

陈英,张少白

(南京邮电大学 计算机学院,江苏 南京 210003)

**摘要:**针对发音信息在语音环境中并不容易得到的问题,提出了一种从听觉信号中预测发音信息的语音反演方法。论文应用远端监督学习(DSL),对语音反演机器学习策略进行研究,并对其实验背景和理论依据进行了分析。论文在提出一种对远端监督学习逆模进行全局优化的方法的同时,通过应用八个声道变量作为发音信息来模拟语音动力学,对语音信号分别被参数化为声学参数(APs)和梅尔频率倒谱系数(MFCCs)时的预测结果进行了比较。结果表明远端监督学习对声道变量有较好的预测性能。

**关键词:**发音信息;语音反演;远端监督学习;声道变量

中图分类号:TP181

文献标识码:A

文章编号:1673-629X(2013)03-0105-04

doi:10.3969/j.issn.1673-629X.2013.03.027

## Research on Distal Supervised Learning Model of Speech Inversion

CHEN Ying,ZHANG Shao-bai

(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

**Abstract:** To the problem that articulatory information is not readily available in typical speaker-listener situations, a method that estimates articulatory information from the acoustic signal is proposed, namely speech inversion. It selects distal supervised learning (DSL) as one of machine learning strategies for speech inversion to study, and analyzes the experiment's background and theoretical foundation of distal supervised learning. It proposes that use a global optimization approach for the inverse model of distal supervised learning and eight tract variables as articulatory information to simulate speech dynamics, the results when speech signal is parameterized as acoustic parameters (APs) and as mel-frequency cepstral coefficients (MFCCs) are compared in the paper. The results show that distal supervised learning has a good estimation performance for tract variables.

**Key words:** articulatory information; speech inversion; distal supervised learning (DSL); tract variables

## 0 引言

目前,自动语音识别系统的性能在非正式的或自发性的语音中会受到影响,这是因为自发性语音有很大的可变性,而这些可变性主要是由协同发音引起的。许多研究提出利用发音信息可以提高自动语音识别系统的性能。可惜的是,这些发音信息在某些特定的对话环境中并不那么容易得到。因此,需要通过一种方法来预测这些信息,这种方法通常称为“语音反演”。

声道器官生成语音信号的过程可以用函数 $f$ 表示,即 $f:t \rightarrow x$ 。式中 $x$ 表示语音信号向量, $t$ 表示发音器官配置向量, $f$ 函数定义从发音域到声音域的前向映射。因此,假设 $f$ 已知,如果给出表示特定发音器官

配置的向量 $t_a$ ,就能得到一个特定的语音输出 $x_a$ 。在语音识别过程中,已知的是语音信号,并没有发音器官的数据。但是如果定义一个函数 $g$ ,即 $g:x \rightarrow t$ ,则发音器官的配置 $t_b$ 可以通过使用函数 $g$ 从语音样本 $x_b$ 得到。因此, $g$ 是 $f$ 的逆函数,表示从声音到发音器官的语音反演。

一些机器学习方法已被用于执行语音反演的任务,如人工神经网络,支持向量回归和自回归人工神经网络。人工神经网络在非线性<sup>[1]</sup>回归问题中用途广泛,但是它不适用于不适定的回归问题,其中不适定性是由一到多的映射引起的;支持向量回归估计发音轨迹时容易受到噪声的影响,所以预测性能不高;自回归人工神经网络的反馈环虽然有助于维持发音轨迹的平滑,但是同时也可能引进累进误差并且计算成本高。而有远端教师监督的学习模型(也可称为远端监督学习,DSL)不仅能解决一到多的映射问题,而且预测性能比人工神经网络、支持向量回归以及自回归人工神经网络好。文中接下来部分主要研究这一模型。

收稿日期:2012-06-06;修回日期:2012-09-11

基金项目:国家自然科学基金资助项目(61073115)

作者简介:陈英(1987-),女,硕士研究生,研究方向为人工智能和模式识别;张少白,硕士生导师,教授,博士,主要研究方向为人工智能与认知科学、信息获取、处理与识别。

## 1 声道变量

文中应用八声道变量<sup>[2,3]</sup>作为发音信息来模拟语音动力学。这八个声道变量描述声道中特定发音器官的收缩程度和位置,分别为唇孔(Lip aperture, LA),唇突(Lip protrusion, LP),舌尖收缩程度(Tongue tip constriction degree, TTCD),舌尖收缩位置(Tongue tip constriction location, TTCL),舌体收缩程度(Tongue body constriction degree, TB CD),舌体收缩位置(Tongue body constriction location, TBCL),软腭(Velum, VEL),声门(Glottis, GLO),如图1所示:

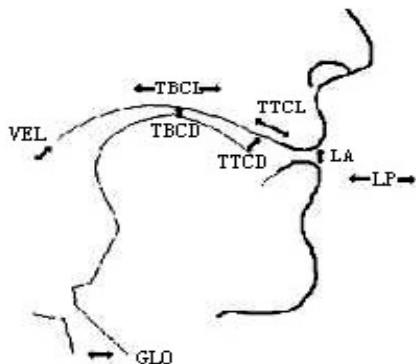


图1 不同收缩位置的声道变量

声道变量<sup>[4-6]</sup>从收缩的程度和位置方面,描述了声道形状的几何特征。一个有效的发音动作由激活开始和结束的时间以及一些参数值来详细说明。这些参数值适用于一组临界阻尼的二阶微分方程<sup>[7]</sup>,如式(1)所示。其中, $M$ 、 $B$ 和 $K$ 分别表示质量,阻尼系数和每个声道变量(用 $z$ 表示)的抗扰性参数,而 $z_0$ 是动作的目标位置。

$$M\ddot{z} + B\dot{z} + K(z - z_0) = 0 \quad (1)$$

使用声道变量有以下三个优点:

- 1) 声道变量直接详细地说明了声道区域函数的特征。
- 2) 声道变量的收缩直接由发音动作控制并体现了讲话者的语音目标。另外,声道变量到语音的映射近乎为一对一的关系,能减少语音反演中的非唯一性。因此从语音分类的方面来说,声道变量能提供更多的信息。
- 3) 结合声道变量信息不仅能提高识别发音动作的性能,而且能改进语音识别系统的鲁棒性<sup>[8]</sup>。

## 2 数据集和信号参数化

研究中使用的数据库主要来自文献[9]。这个数据库使用 TADA 模型和 Hlsyn 语音合成器来生成,数据库包含了合成的语音以及它们的发音说明。合成数据库通过输入 420 个不同单词的文本产生。输出的合成语音以 10kHz 进行采样,而声道变量时间函数和动

作得分则以 200Hz 进行采样。75% 的数据用于训练,10% 的数据用于验证,剩下的数据用于测试。

语音信号被参数化为声学参数<sup>[10,11]</sup>(AP)和梅尔倒谱系数(MFCC)。根据声学参数的相关性,挑选出 40 个不同的声学参数,而梅尔倒谱系数则提取 13 个。声学参数用长度为 10ms,帧频为 5ms 的窗口进行测量。梅尔倒谱系数的声学特征也是以长度为 10ms,帧频为 5ms 的窗口(与声道变量的时间同步)进行测量。将声学特征和目标发音信息(即声道变量)进行 $z$ 归一化,并进行缩放,使得它们的动态范围限制在 $[-0.95, +0.95]$ 。根据已有的观察可知,结合动力学信息有助于减少语音反演任务的非唯一性问题。因此,文中的实验都是将输入特征置于上下文中研究。特征的语境化用上下文窗口参数 $\hat{C}$ 定义。而当前帧(特征范围为 $d$ )与之前或之后的 $\hat{C}$ 帧(帧偏移 2 个或时间偏移 10ms)连接起来,连接成的特征向量大小为 $(2\hat{C} + 1)d$ 。根据已有的研究<sup>[12]</sup>可知,对于梅尔倒谱系数来说最佳的上下文参数 $\hat{C}$ 值是 8(上下文为 170ms),而对于声学参数来说最佳上下文参数值是 9(上下文为 190ms),论文接下来描述的实验部分就是使用此值。

## 3 远端监督学习

为了解决常见的监督学习结构在一到多映射情况中的问题,Jordan<sup>[13]</sup>等人提出了有远端教师的监督学习(或称为远端监督学习)。

在远端监督学习模型中,有两个模型相串联:

- 1) 前向模型(产生给定发音轨迹的发音特征,即 $M$ 到 1 的映射);
- 2) 逆模(根据发音特征产生发音轨迹,即 1 到 $M$ 的映射)。

给定一组 $[x_b, y_b]$ 的数据对,远端监督学习首先对前向模型进行学习,结果是唯一的但不一定是完美的。远端监督学习通过将逆模与前向模型串联从而对逆模进行学习,如图2所示。远端监督学习结构可以理解作为一种“综合分析”的方法,前向模型是综合阶段,逆模是分析阶段。在远端监督学习方法中,前向模型的权值和偏差保持常量,而逆模的误差通过前向模型反向传播给逆模,这样逆模利用这些误差进行训练,而它的权值和偏差也会随之更新。

考虑到网络权值向量 $w$ 和偏差向量 $b$ ,则输入向量 $x$ 和输出向量 $y$ 间的前向映射关系可表示为:

$$\hat{t} = g(x, w, b) \quad (2)$$

基于下面的代价函数对前向模型进行学习:

$$L = \frac{1}{2} E[(t - \hat{t})^T (t - \hat{t})] \quad (3)$$

其中 $t$ 是对于给定输入的期望目标。

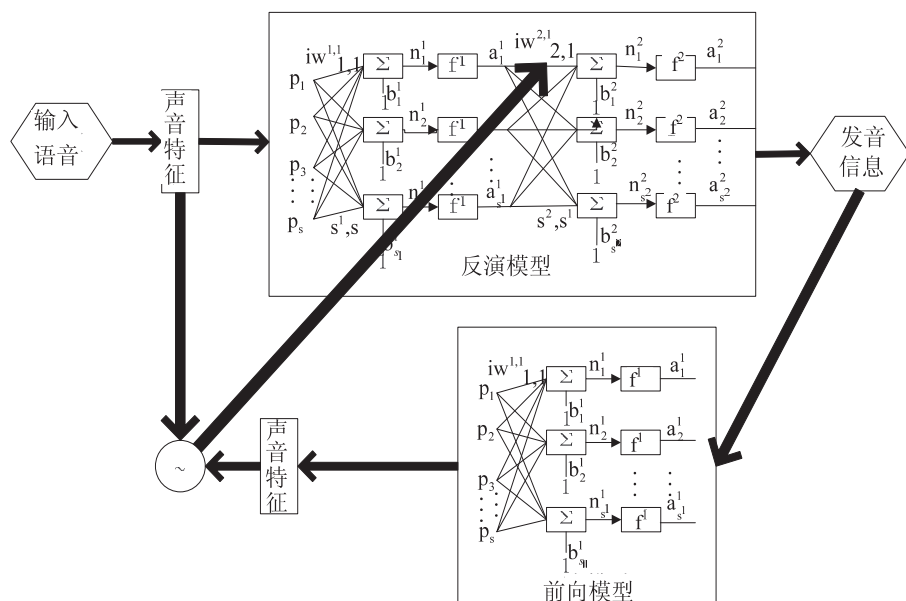


图2 用于获取声音到声道变量映射的  
远端监督学习模型

对于逆模,文献[13]定义了两种不同的方法,一种是局部优化方法,另一种是根据轨迹优化的方法。局部优化方法必须使用在线学习规则,而根据轨迹优化的方法要求在网络中循环(及时使用反向传播能使误差最小),但两种方法都明显增加了训练时间和内存需求。文中使用一种全局优化的方法,这个方法与局部优化方法类似,也是使用远端监督学习工具,但在前馈网络中进行批训练,这就能大大减少训练时间。

远端监督学习的代价函数(力求最小)可表示为:

$$J = \frac{1}{2N} \sum_{k=1}^N [(\mathbf{t}_k^* - \mathbf{t}_k)^T (\mathbf{t}_k^* - \mathbf{t}_k)] \quad (4)$$

其中  $N$  是训练样本的总数,  $\mathbf{t}_k$  是第  $k$  个训练样本的目标向量,  $\mathbf{t}_k^*$  是网络的实际目标输出。

权值更新规则如下:

$$w[n+1] = w[n] - \eta \nabla_w J_n \quad (5)$$

其中  $\eta$  是学习速率,  $w[n]$  表示时间序列  $n$  下网络的权值。

梯度则可以通过将式(4)代入到下面的链式法则计算得到:

$$\nabla_w J_n = \frac{1}{N} \sum_{k=1}^N \left( -\frac{\partial \mathbf{x}_k^T}{\partial \mathbf{w}} \frac{\partial \mathbf{t}_{k,n}^*}{\partial \mathbf{x}_k} (\mathbf{t}_k - \mathbf{t}_{k,n}^*) \right) \quad (6)$$

其中  $\mathbf{t}_{k,n}^*$  是第  $k$  个训练样本在第  $n$  时刻的预测目标向量。

## 4 实验,结果和讨论

远端监督学习结构对八个声道变量轨迹的每个声学特征都进行训练。前向模型用单层隐层的前馈人工神经网络创建,并用尺度共轭梯度算法训练。隐层中

的神经元数通过在确定集上使用均方根误差来优化。逆模用三层隐层的网络建立,而每层隐层的神经元数通过在确定集上使用均方根误差来优化。远端监督学习模型使用梯度下降算法(以可变的学习速率)和动量学习准则(动量=0.9)进行训练。前向模型的神经元数为300和400,逆模中对应梅尔倒谱系数和声学参数的神经元数分别为150-100-150和250-300-250。

实验想要演示的是给定语音信号,高度准确地预测出声道变量。实验使用两种定量测量方法,将预测的发音轨迹的形状和动态与实际的进行比较,这两种量分别是:均方根误差(RMSE)和皮尔森积差相关(PPMC)系数。均方根误差表明了实际发音轨迹与预测发音轨迹整体的区别,而皮尔森积差相关可以测量两者间的振幅和动态相似性。

均方根误差和皮尔森积差相关的定义如下:

$$\text{RMSE} = \sqrt{\frac{1}{N} (\mathbf{e} - \mathbf{t})^T (\mathbf{e} - \mathbf{t})} \quad (7)$$

$$\gamma_{\text{PPMC}} = \frac{N \sum_{i=1}^N e_i t_i - \left[ \sum_{i=1}^N e_i \right] \left[ \sum_{i=1}^N t_i \right]}{\sqrt{N \sum_{i=1}^N e_i^2 - \left( \sum_{i=1}^N e_i \right)^2} \sqrt{N \sum_{i=1}^N t_i^2 - \left( \sum_{i=1}^N t_i \right)^2}} \quad (8)$$

其中  $\mathbf{e}$  表示预测的声道变量向量,  $\mathbf{t}$  表示有  $N$  个数据点的实际声道变量向量,  $N$  是考虑到的发音轨迹的个数,声道变量中个数是8。

从远端监督学习模型中得到的声道变量的预测结果如图3、图4所示,语音信号分别被参数化为声学参数和梅尔倒谱系数。

从图3可以看出,将语音信号参数化为梅尔倒谱系数与参数化为声学参数相比,前者六个声道变量(除软腭和唇突)的皮尔森积差相关值较大。从图4可以看出,将语音信号参数化为梅尔倒谱系数与参数化为声学参数相比,前者六个声道变量(除软腭和唇突)的均方根误差较小。值得注意的是,较低的均方根误差和较高的皮尔森积差相关表示预测的性能较好。因此从整体来看,声学参数对软腭和唇突的预测有较高的准确性,然而对其它声道变量的预测,梅尔倒谱系数有较好的结果。从图3、图4还可以看出,在对八个声道变量的预测中,远端监督学习对声门的预测

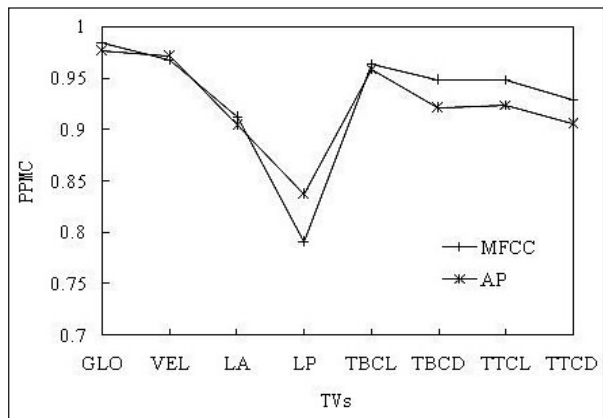


图 3 参数化为梅尔倒谱系数和声学参数时对声道变量预测的皮尔森积差相关

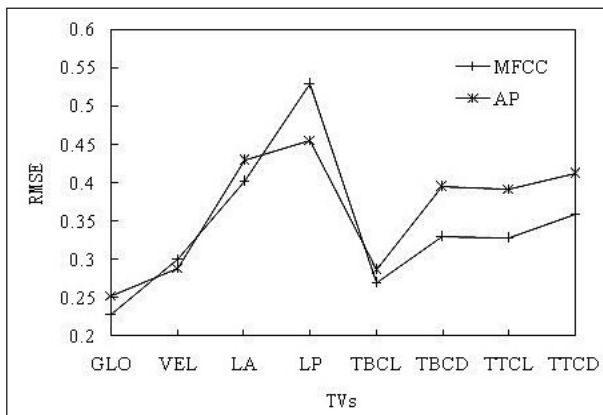


图 4 参数化为梅尔倒谱系数和声学参数时对声道变量预测的均方根误差

性能最好,对唇突的预测性能最差。总体来说,大部分声道变量都有较高的皮尔森积差相关值和较低的均方根误差值,因此可以说远端监督学习对声道变量有着很好的预测性能。

## 5 结束语

在笛卡尔坐标系上,多个不同的发音位置有可能表示相同的声道收缩,而声道变量的表示则是唯一的。McGowan 也已指出使用声道变量可以改善语音反演中的非唯一性。另外,前面已经指出,远端监督学习能克服人工神经网络在不适应回归问题中的缺点,并且比支持向量回归、自回归人工神经网络的预测性能更好,因此可以说远端监督学习在语音反演中有很重要的作用。不过,正如前面指出的,远端监督学习拓扑非常像综合分析结构,其中综合部分的性能完全依赖于前馈模型的准确性。为了保证前馈模型高度的准确性,需要详细的数据来保证前馈模型有所有可能的发音数据和语音观测的数据对,这是远端监督学习的缺点所在,但是对其应用价值影响不大。

## 参考文献:

- [1] Neiberg D, Ananthakrishnan G, Engwall O. The acoustic to articulation mapping: non-linear or non-unique [C]//Proc. Interspeech, 9th Annual Conference of the International Speech Communication Association. Australia: [s. n.], 2008: 1485-1488.
- [2] Zhuang X, Nam H, Hasegawa-Johnson M, et al. The entropy of articulatory phonological code: recognizing gestures from tract variables [C]//Proc. Interspeech, 9th Annual Conference of the International Speech Communication Association. Australia: [s. n.], 2008: 1489-1492.
- [3] Zhuang X, Nam H, Hasegawa-Johnson M, et al. Articulatory phonological code for word classification [C]//Proc. Interspeech, 10th Annual Conference of the International Speech Communication Association. U. K.: [s. n.], 2009: 2763-2766.
- [4] Mitra V, Nam H, Espy-Wilson C, et al. Retrieving tract variables from acoustics: a comparison of different machine learning strategies [J]. IEEE Journal of Selected Topics in Signal Processing, 2010, 4(6): 1027-1045.
- [5] Katsamanis A, Papandreou G, Maragos P. Face active appearance modeling and speech acoustic information to recover articulation [J]. IEEE Trans. on Audio, Speech, Lang. Process., 2009, 17(3): 411-422.
- [6] Mitra V, Özbek I, Nam H, et al. From acoustics to vocal tract time functions [C]//Proc. of ICASSP. [s. l.]: [s. n.], 2009: 4497-4500.
- [7] Byrd D, Saltzman E. The elastic phrase: modeling the dynamics of boundary-adjacent lengthening [J]. J. Phonetics, 2003, 31(2): 149-180.
- [8] Mitra V, Nam H, Espy-Wilson C, et al. Noise robustness of tract variables and their application to speech recognition [C]//Proc. Interspeech, 10th Annual Conference of the International Speech Communication Association. U. K.: [s. n.], 2009: 2759-2762.
- [9] Nam H, Goldstein L, Saltzman E, et al. Tada: an enhanced, portable task dynamics model in matlab [J]. J. Acoust. Soc. Amer., 2004, 115(5-2): 2430-2430.
- [10] Juneja A. Speech Recognition Based on Phonetic Features and Acoustic Landmarks [D]. USA: Univ. of MD, College Park, 2004.
- [11] He X, Deng L. Discriminative learning for speech processing [C]. CA: Morgan & Claypool, 2008.
- [12] Mitra V, Nam H, Espy-Wilson C. A step in the realization of a speech recognition system based on gestural phonology and landmarks [J]. J. Acoust. Soc. Amer., 2009, 125(4): 2530-2530.
- [13] Jordan M I, Rumelhart D E. Forward models-supervised learning with a distal teacher [J]. Cogn. Sci., 1992, 16: 307-354.

# 语音反演远端监督学习模型研究

作者: [陈英, 张少白](#)  
作者单位: [南京邮电大学 计算机学院, 江苏 南京210003](#)  
刊名: [计算机技术与发展](#)  
英文刊名: [Computer Technology and Development](#)  
年, 卷(期): 2013(3)

本文链接: [http://d.g.wanfangdata.com.cn/Periodical\\_wjtz201303029.aspx](http://d.g.wanfangdata.com.cn/Periodical_wjtz201303029.aspx)