# 一种面向混合语言的语音合成方法

## 背景介绍

语音合成技术是将文本转化成声音的技术。历史上语音合成技术经过规则合成、拼接合成、统计概率模型合成三个阶段，当前新出现的方法是基于神经网络的合成方法。在这一方法中，神经网络用作映射函数，将输入的文本信息转换成基频、频谱等发音参数。

混合语言语音合成是指待合成文本中存在多种语言。这种混合语言语音合成一向是技术难点，**一个重要原因是数据库中多语言发音者发音很不一样（找到一个会发各种语言的发音者几乎是不可能的），这导致从一种语言跨越到另一种语言时会产生显著的变声。**在统计模型时代，有可能的解决方法包括：

* 模型自适应。例如语言A的发音者是m ,语言B的发音者是n，二者单独训练声学模型MAm和MBn，但m也可以发少量B语言的声音，因此可利用m在B语言上的发音对MBn做自适应（如MAP或MLLR），得到MBnm，再将MAm和MBnm做混合语言发音模型。这一方法的缺点是必须有会说多种语言的发音人，而且自适应在句子数较少时并不得取得听起来非常接近的效果。
* 模型映射。另一种解决混合语言发音的方法是模型映射法。同样，让发音者m和n分别训练本语言的模型MAm和MBn，考虑到不同语言其基础发音是十分相似的，只不过具体拼接起来有所不同。这种“原子发音”的相似必可以用来实现模型映射。

例如我们现在想让m的声音发B语言，而我们只有发A语言的模型。怎么办呢？我们可以假设让n的声音发B语言，在发音空间中有一条n发B语言应选择哪些“原子发音”的路径，将这条路径映射里MAm模型里的路径，再利用MAm进行发音，听起来就象是m在发B语言。这里的“原子发音”是概率方法里隐马尔可夫模型的状态，或称seno. 这一方法在拼接模型里也适用，只要找到相似的发音单元即可。这一方法的好处是模型可以单独训练，不需要发音人发多种语言，混合起来比较自然，缺点在于合理的映射并不好找，拼出来的声音也会显得带有带有本族语口语，表现不自然。

## 发明内容和思路

本发明提出一种基于神经网络的混合语言语音合成方法，其基本思路是，用**多语言多发音人数据**混合语言发音模型，但在**训练时将将发音人信息从发音信号中剥离**。这相当于对信号做了**面向发音人的正规化**，基于这种正规化后的神经网络模型**仅学习发音内容**，在实际合成时再把发音人信息加入。基于这种方法，不仅可以**让同一发音人发多种语言的声音**，而且可以任意**改变发音人特性**，得到个性化的语音合成系统。

## 发明要点

本发明包括如下三个部分：说话人特征提取，基于说话人正规化的多语言数据神经网络模型训练，基于说话人特征向量的多语言发音。

1. **说话人特征提取**

说话人特征提取可采用多种模型，包括于i-vector模型，CNN或RNN模型。说话人特征归结为一个向量表示，我们称为speaker vector，或s-vector。对训练数据中的所有说话人提取s-vector模型，每个说话人的每句话的s-vector相同，且需通过LDA将语言、信道等信息滤除。

1. **基于说话人正规化的多语言数据神经网络模型训练**

本发明的关键在于利用多发音人、多语言数据进行混合语言混练。我们需要训练的模型采用递归神经网络（ＲＮＮ），其结构如下：输入为两组：**一组说话人特征向量s-vector，一组由发音文本生成的语言学向量**（如上下文音素、是否词边界、是否语言边界、音素在词中的位置等）。通过加入s-vector，训练即达到对说话人正规化的效果。输出为三组预测值：1维基频，1维非周期激励，若干维频谱。训练数据包括多种语言和多个发音人的数据。和传统拼接方法不同，我们的方法允许利用同一语言的不同发音人，不同语言的不同发音人数据同时训练，由于有s-vector做规化，生成的模型将剥离发音人的属性，仅关注由语言学特征到声学层的映射。

混合语言发音模型

声学模型

语言特征提取

**英语数据**

**汉语数据**

i-vector CNN/RNN

**s-vecotr**

**日语数据**

**图1： 模型训练过程**

1. **基于说话人特征向量的多语言发音**

模型训练完毕后，该模型就有了接收一个说话人特征向量s-vector，生成符合该说话人特征的混合语言发音的能力。S-vecotr可由某一发音人的发音数据生成（如某个训练集中的发音者，或希望听到的发音人），也可以人为自由调整，生成具有个性的声音。

语音信号生成

**s-vecotr**

声学模型

语言学特征提取

**发音文本**

**图2. 基于s-vector的多语言语音合成**

## 发明优势

1. 不需同一发音人的多语言数据，实现自然连续的多语言混合发音。
2. 可实现对发音人特性的自由修改。