CSLT TECHNICAL REPORT-20170001 [Friday 30th June, 2017]

重叠语音与原始语音关系的研究与分析

Hui Tang

Correspondence:
tanghui@cslt.riit.tsinghua.edu.cn
Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China
Full list of author information is available at the end of the article

Abstract

在公共场所中，我们的声音不可避免的和外界的各种声音交织在一起，比如音乐声，别人的说话声等等。但是我们却能在注意力集中在某个人的谈话之中而忽略背景中其他的对话和噪音。这就是著名的鸡尾酒效应。即我们可以在所有人的声音中听懂某个人的声音。但是怎么让机器在这种情况下“听懂”某个人的声音，仍然是一个悬而未决的问题。同时我们把多个人同时说话的语音称为重叠语音，这里的每个人自己说的话称为原始语音。本文主要是研究两个方面，一是检测一句话中是否存在重叠语音，二是重叠的语音与原始语音的关系，以便为后续研究做铺垫。

Keywords: 语音重叠; 检测语音重叠; 原始语音与重叠语音的关系

1 介绍

在日常生活中，我们大部分的交流处于在有噪声的环境中。其实对于有噪声的语音识别，现在的技术已经做的非常的成熟了，但是对于怎么识别重叠语音，仍然是一个非常困难的问题，尤其是怎么分离出重叠语音中的原始语音进而分别识别他们。带着这个问题，我们开始研究重叠的语音与原始语音的关系。期望能够通过它们之间的关系帮助我们从重叠语音中分离出每个人说的语音。本文通过实验，将专门讨论和研究这两者之间的关系。

2 实验1-检查语音是否是重叠的

2.1 数据集

该实验的训练集是th30，由两个部分组成：其中原始数据是th30的训练集（10000 句），重叠语音是使用th30 中训练集的语音用sox 指令把不同的两句话合成的（10000 句）。

另外测试集是两个：一是th30的测试集和使用该测试集合成的语音（共4990句）；第二个是真实环境下录的，包含不重叠的声音（7人，每人3句话），两个人齐声朗读一句话（4组，每组3 句），两个人说的不同的话（6组，每组3句）。语音在/work7/tanghui/overlap/data_real_envir下

2.2 模型训练

训练的模型就DNN，共输入9帧（从前4帧到后4帧）。将上面的训练集进行特征提取，然后对原始语音的特征做vad，重叠语音特征的vad是原始语音的vad 的交集，然后分别根据vad去除静音，得到的结果作为dnn的输入。然后再将做完vad之后的原始语音的特征按帧标记为0，同理，将做完vad之后的重叠语音的特征按帧标记为1，然后作为模型的输出。（脚本：/work7/tanghui/overlap/run_tdnn_spk_pure.sh）

2.3 测试结果

测试集	原始语音错误率	重叠语音错误率	总错误率
th30_1	19.4%	10.7%	15.3%
th30_2	19.4%	15.3%	17.3%

测试集	原始语音错误率	说相同的话的错误率	说不同的话的错误率
真实环境	19.4%	28%	37.8%

th30_1指的是重叠语音是原始语音的vad的交集得到的，th30_2指的是重叠语音只是对自身做vad得到的。th30_2用来和下面的实验做对比。

在第二个测试集上发现不同的人的错误率差别很大，有的人40%-60%的错误率，而有的人10%-20%的错误率。（结果:/work7/tanghui/overlap/data_real_envir/test_diff_context/frame_test_diff_context/score.log）

3 实验2-研究原始语音与重叠语音的关系

3.1 数据集

实验的数据集分成三个：两个不同的人说不同的话，同一个人说不同的话，同一个句话，不同的音量。所有的数据都是来自th30，实验在/work7/tanghui/gene_400目录下.

语音降低音量指令：

sox -v 0.1 A2_0.wav 0.1out_A2_0_A2_1.wav

将A2_0.wav语音的音量变为现在的1/10，保存为0.1A2_0.wav

语音先降低声音后再重叠

sox -m -v 0.1 A2_0.wav A2_1.wav 0.1out_A2_0_A2_1.wav

将A2_0.wav的音量将为现在的1/10，然后在和A2_1.wav合并为0.1out_A2_0_A2_1.wav。

以上三个数据集都是th30上合成的。

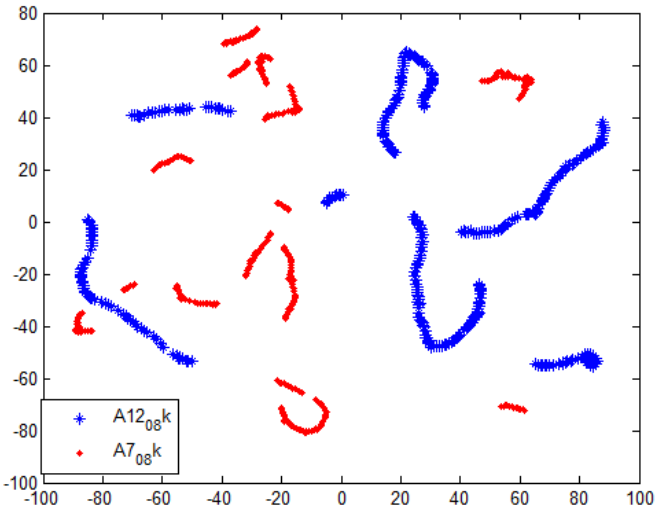
3.2 模型

模型是蓝天哥做speaker factorization的模型，主要是能够在帧级别识别说话人。(模型在/work7/tanghui/gene_400/cnn_4_8_pi2000_po400_l6_splice10.)

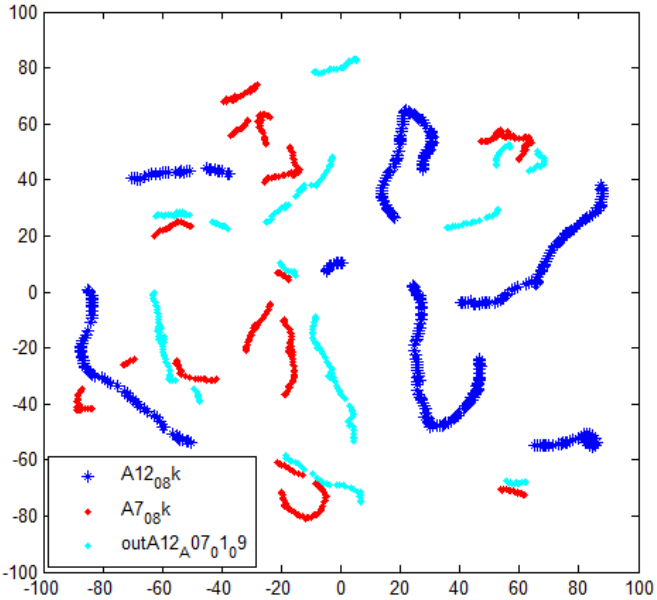
3.3 画图

画图显示使用的是t-sne，一个MATLAB的工具箱，能够将高维数据降到低维，然后将结果画出来。代码在192.168.0.200的服务器上.
（E:/test_th/spk_overlap/8k/run_tsne_plot_one_voice_multi_word.m）

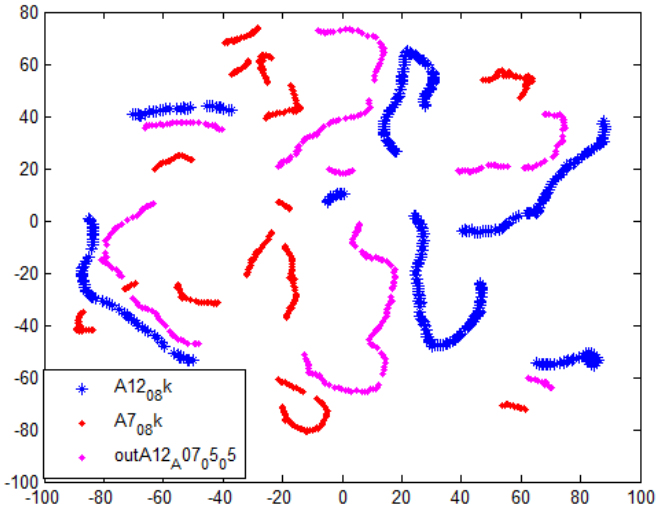
3.4 两个不同的人



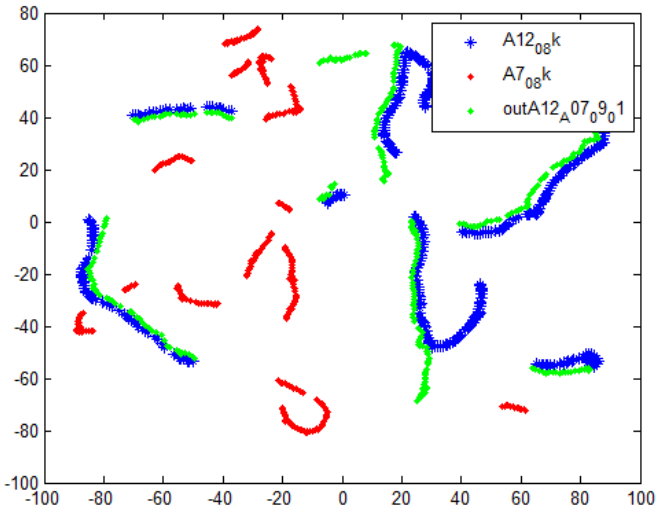
上图是两个不同的人说的两句不同的话，红色是A12，蓝色是A7



红色和蓝色与上图一致,青色是A12的声音变为原来的1/10与A7的音量变为原来的9/10 之后合成的，可以看出青色与A7(红色)更靠近一些。

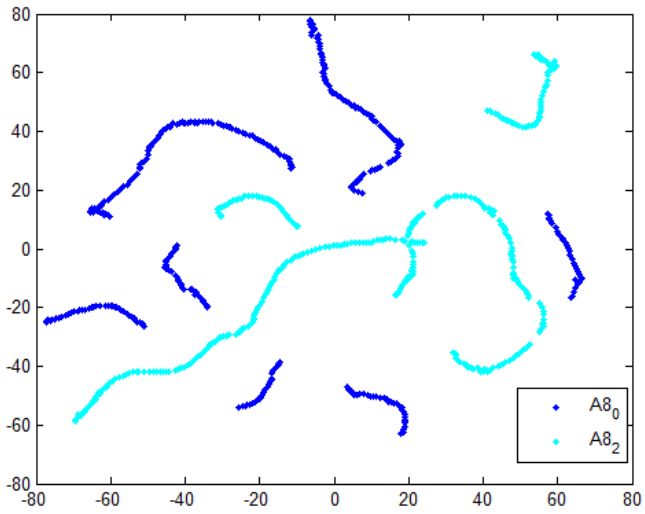


粉红色为A7与A12两个人的音量都将为原来的1/2之后合成的，所以它就在红蓝之间。

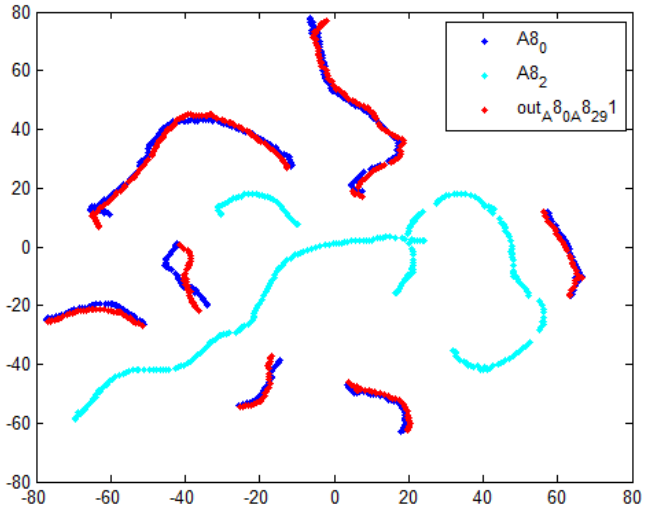


绿色代表A12的音量将为原来的9/10，A7变为1/10合成的，它更接近A12.

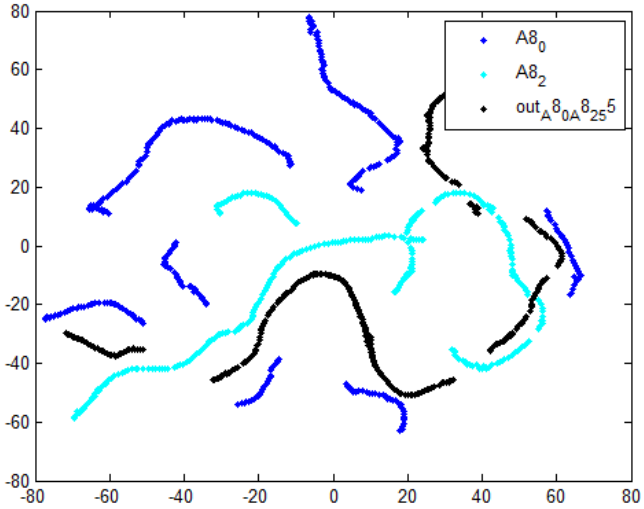
3.5 同一个人，不同的话



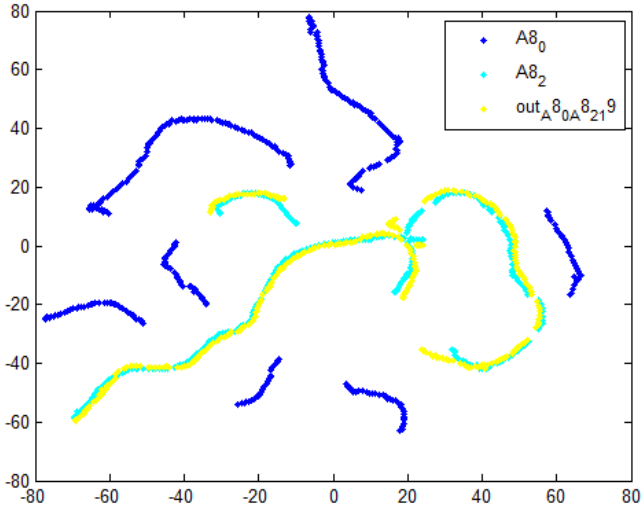
一个人A8说的两句不同的话.



红色表示由蓝色代表的话的音量的9/10与青色代表的话的音量的1/10合并成的。它基本与蓝色的话重合。

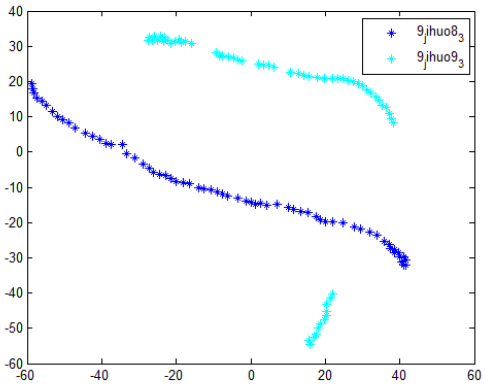


黑色表示蓝色与青色的音量的1/2合成的，它在蓝色和青色之间。

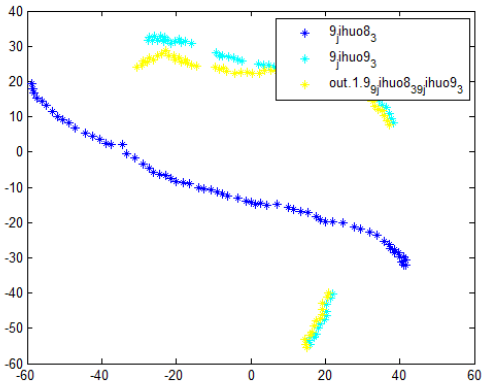


黄色表示用蓝色代表的话的音量的1/10与青色代表的话的音量的9/10合并成的。它基本与青色的话重合。

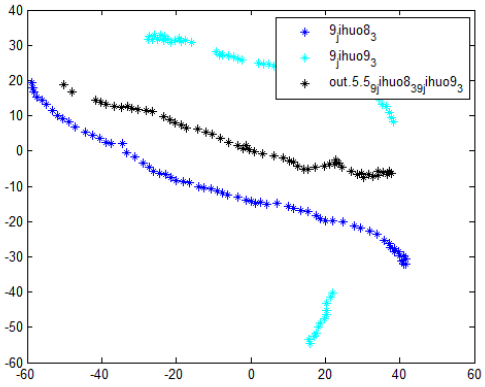
3.6 同一个人，相同的话



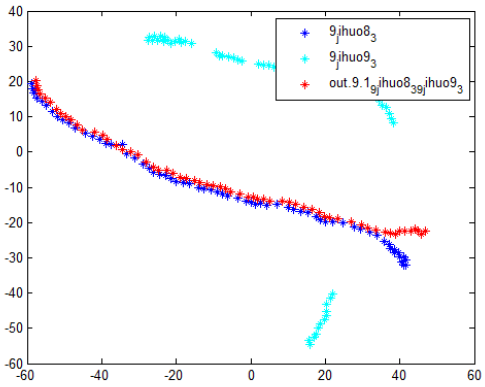
一个人说两句相同的话.



黄色是蓝色声音降为原来的1/10和青色的声音降为原来的9/10合并成的，与青色基本重合。



黑色表示蓝色与青色的音量的1/2合成的，它在蓝色和青色之间。



红色是蓝色声音降为原来的1/10和青色的声音降为原来的9/10合并成的。它基本与蓝色重合。

4 结论

通过以上的两个实验，我们得出以下结论。对于一个实验来讲，用th30训练的检测一句话是否有重叠语音的效果还是比较好的，在我自己录的声音上测试，也可以看出这个模型能够较高的可靠性。但是我们发现，不同的人的声音结果相差很大。这个由于现在数据较少，暂时不能判定具体原因是什么。对于第二个实验，我们发现合成之后的语音更加会倾向于声音大的那个.