重叠语音与原始语音关系的研究与分析

Hui Tang, Dong Wang, Lantian Li, Zhiyuan Tang

摘要：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。本文主要是研究重叠的语音与原始语音的关系，为后续研究做铺垫。

关键词：语音重叠，原始语音，两者关系

1 引言

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 数据处理和模型

本次实验的数据，主要来自 th30，然后使用服务器自带的工具 sox 进行处理。

2.1 语音重叠

```
sox -m [input file1] [input file2] … [out file]
```

比如：

```
sox -m A2_0.wav  A2_1.wav  out_A2_0_A2_1.wav
```

将 A2_0.wav A2_1.wav 合并为 out_A2_0_A2_1.wav

2.2 语音降低音量

```
sox -v [number] [input file] [out file]
```

比如：

```
sox -v 0.1 A2_0.wav 0.1A2_0.wav
```

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

2.3 语音先降低声音后再重叠

```
sox -m [input file1] [input file2] … [out file]
```

比如：

```
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。

更多关于 sox 的指令的信息，可以在服务器上输入 sox 之后查看。

注意:本次实验的都只是使用了两个人的语音进行叠加。

2.4 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。

2.5 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。

3 实验

3.1 不同的两个人的语音进行合并

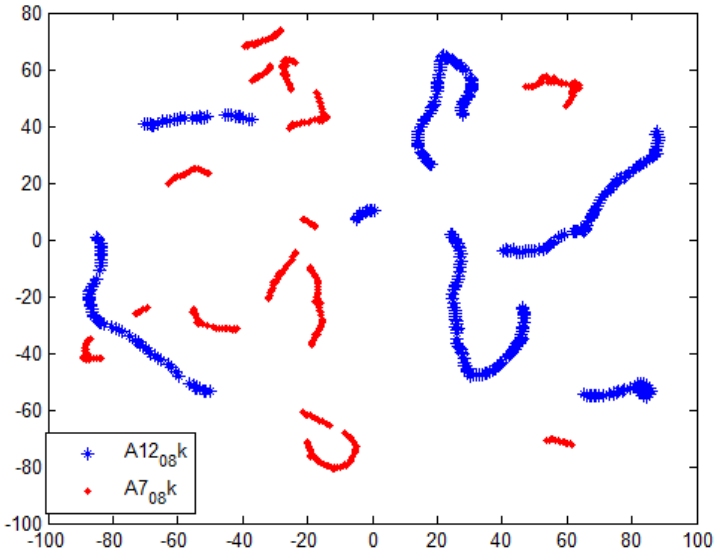

图 1

上图是两个不同的人说的两句不同的话，红色是一个人，蓝色代表另一个人（A12,A7）。

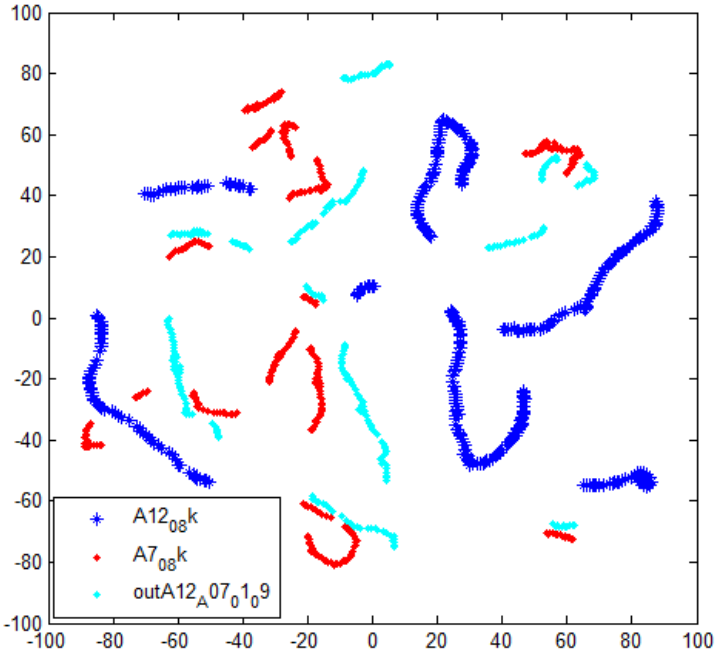

图 2

红色，蓝色分别代表两个人:A12,A7，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

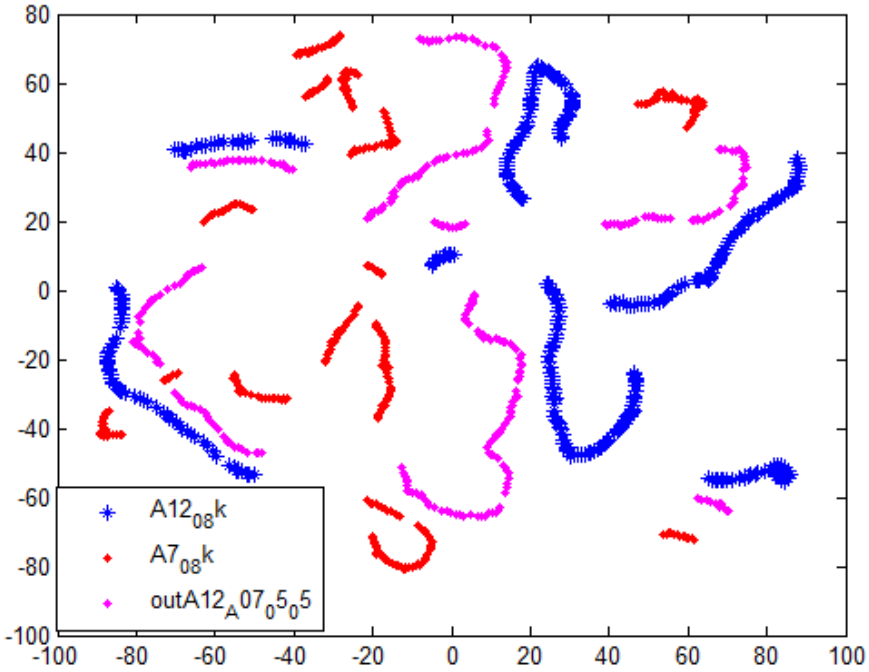

图 3

红色，蓝色分别代表 A7,A12，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。

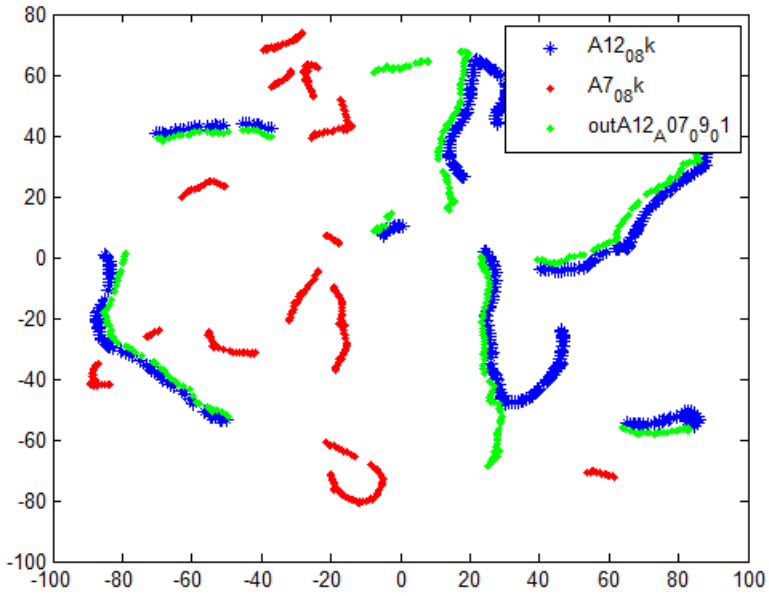

图 4

红色，蓝色分别代表 A7,A12。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.

从以上看出的初步结论就是当不同的人的时候，谁的声音大，合成之后的语音就更偏像谁。

3.2 一个人说的不同的话进行合并

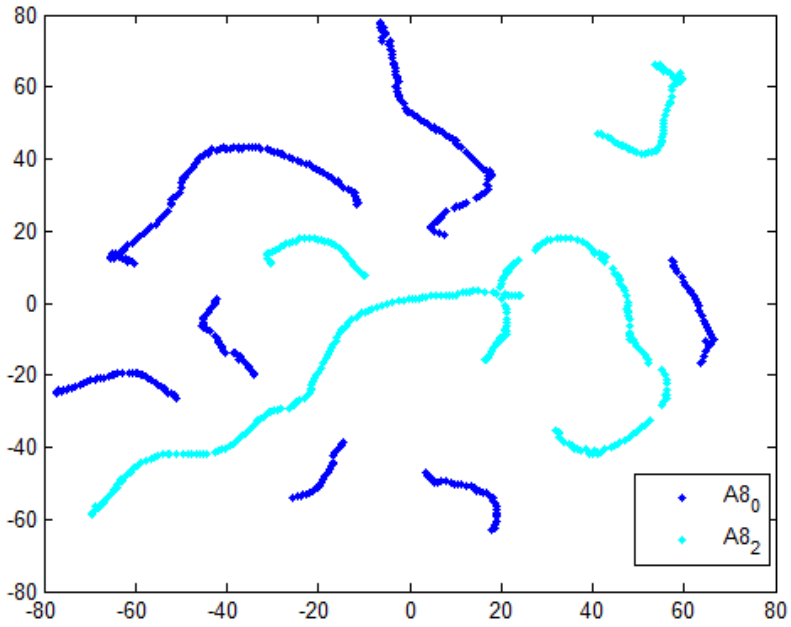

图 5

一个人 A8 说的两句不同的话.

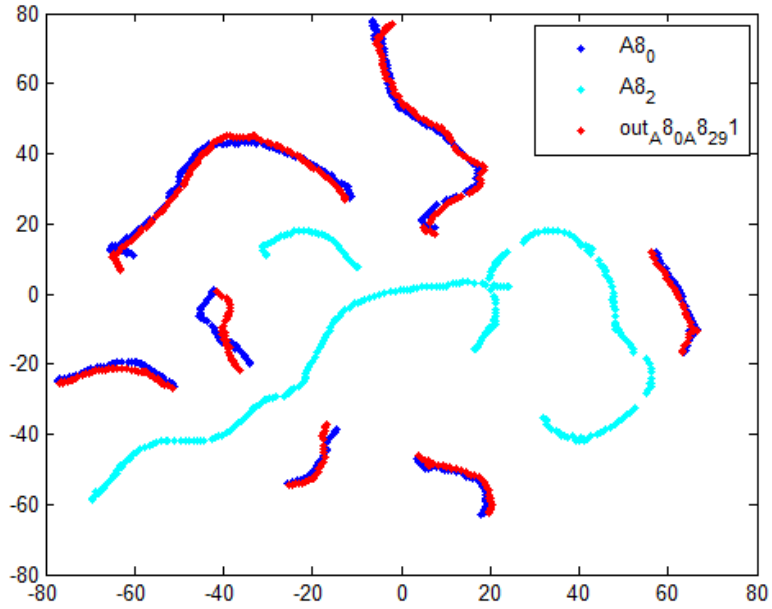

图 6

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。

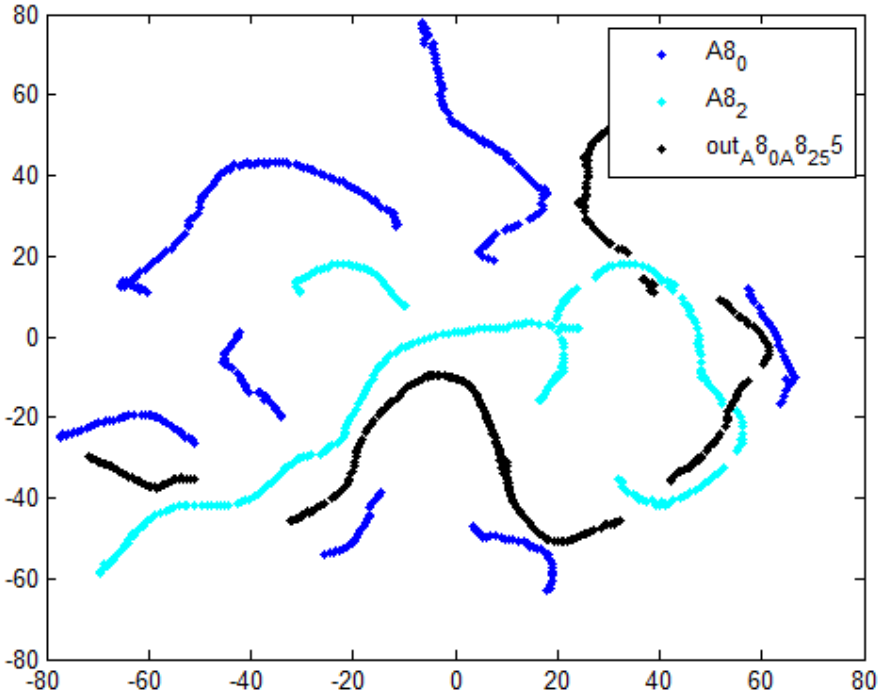

图 7

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。

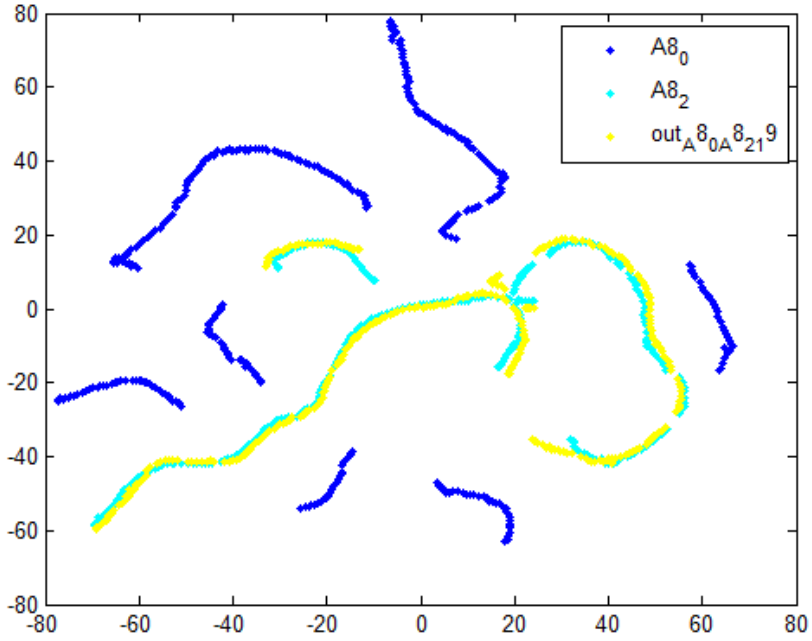

图 8

红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。

这个实验表明把一个人两个的声音的合成之后，也是更倾向于合成之前音量更大的声音。

3.3 一个说两句相同的话

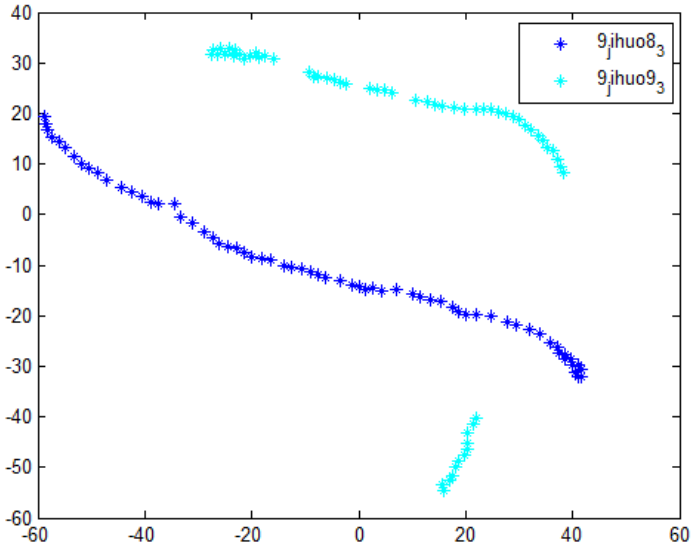

图 9

一个人说两句相同的话

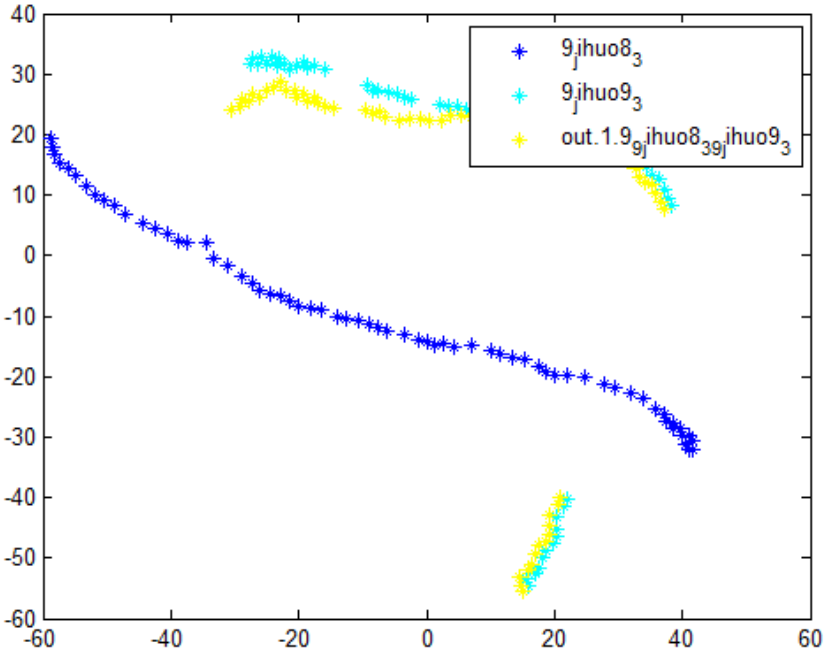

图 10

黄色是蓝色声音降为原来的 9/10 和青色的声音降为原来的 1/10 合并成的。

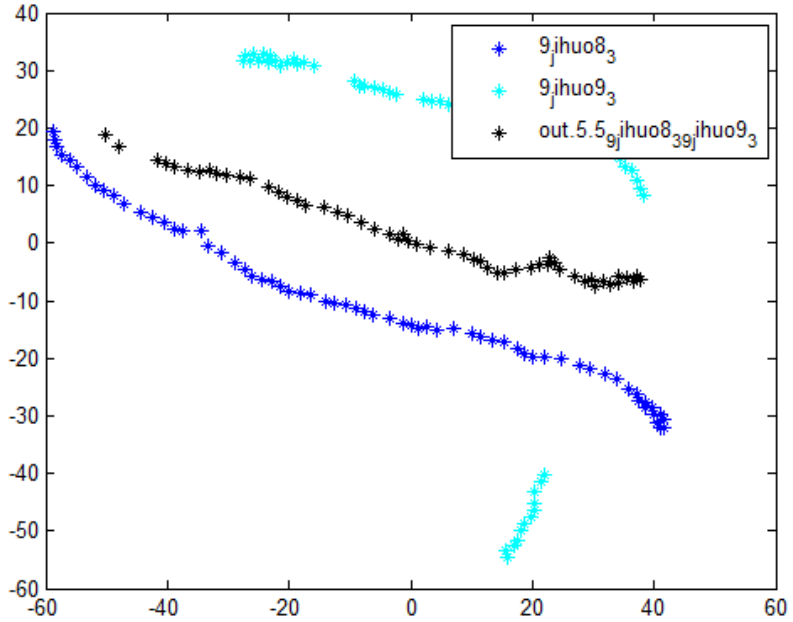

图 11

黑色是青色与蓝色都降为原来的音量的 1/2 合成的。

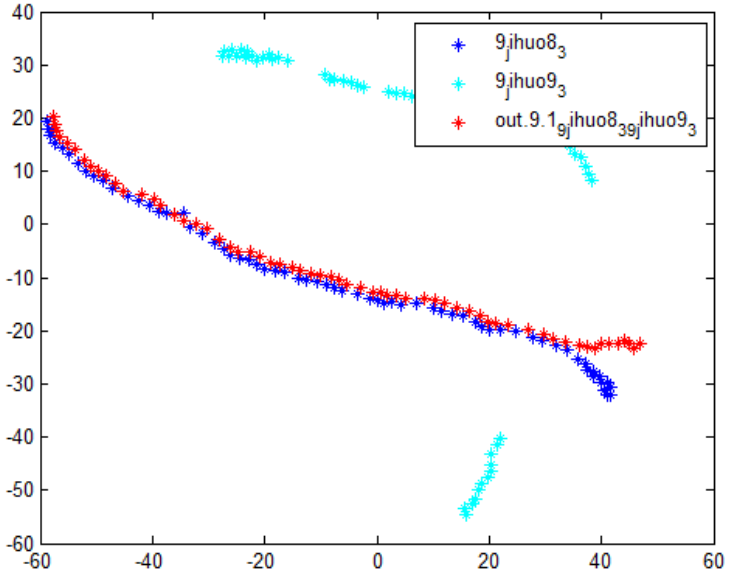

图 12

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。

3.4 同一句话，不同的音量

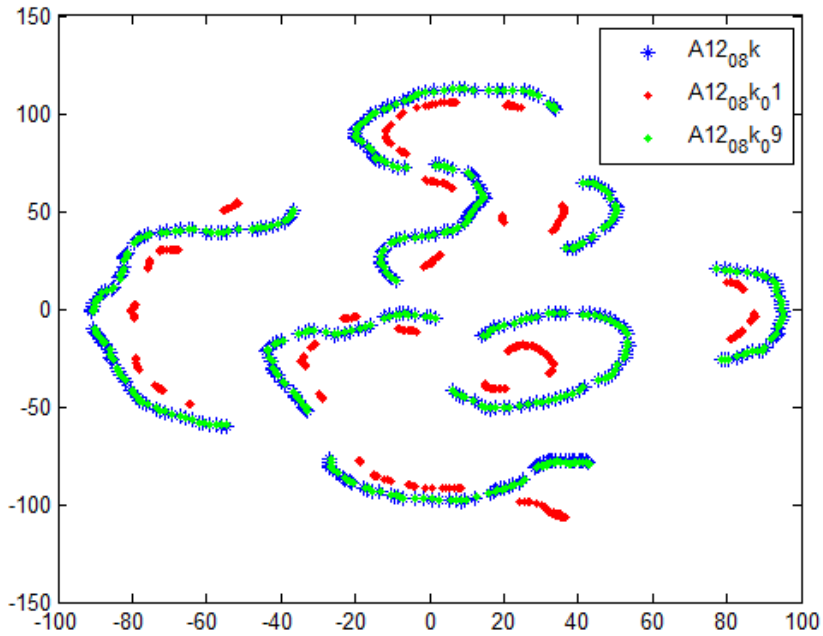

图 13

蓝色原始语音，青色为原始语音的 9/10，红色为原始语音声音的 1/10，从中我们可以得出结论声音越小，离原始声音越远。否则声音越大就会越近。