

Your company just developed a speech recognizer, PERFECT, that is 100% correct. The CEO of the company, Show Me the Money (SMM), wants to know what are the other possible uses for PERFECT. Being a high-strung, creative, student you suggest that the transcripts might be useful for speaker recognition by looking for word idiolect. The CEO asks you to investigate this possibility.

Knowing that you don't have much time you select Switchboard 1 as the corpus for evaluation. The first thing you do is process all of the Switchboard 1 data using an acoustic-based Gaussian Mixture Model – Universal Background Model (GMM-UBM), the current state-of-the-art speaker recognizer. Then you focused on those speakers where the GMM-UBM performed poorly. You noted that one of the speakers, Jack, performed extremely bad.

You process all of Jack's data using PERFECT along with a number of other conversations from other speakers. Next you build a bigram model using all but one of Jack's transcripts. Your hope is to be able to find Jack's transcript among the other speaker's transcripts.

Provided in the Table below is a subset of the bigram model derived from Jack's transcripts. The bigrams are listed in alphabetical order with their number of occurrence. Your objective is to find the bigram that provides the most speaker recognition power via word idiolect.

Table: Jack's Bigram Model				
Bigrams	Count		Bigrams	Count
and and	19		is that	12
and i	18		is uh	15
and uh	32		kind of	24
and the	13		like that	12
and then	9		lot of	15
but uh	8		of the	27
do you	9		on the	9
going to	9		one of	12
have to	15		say it	29
how shall	20		shall i	20
how should	11		that i	29
i don't	17		the other	13
i guess	13		they have	12
i i	14		to be	14
i say	42		you know	48
in the	26			

Questions:

1. Where do you start? What logic did you use to select a starting point?
2. Which conversation(s) match the training data? Which bigrams did you use to determine this?
3. Is there a longer n-gram that would have been more useful or were bigrams effective enough?