Evaluation of the practicality of brute force Speaker Identification in massive sets of calls

Proposal of an alternative approach that is more practical and provides an improved benefit:cost ratio

Content

Executive summary

Two alternatives have been considered to the problem of finding the call of a known target in a large number of calls:

- **Alternative A:** Applying brute force Speaker Identification on the traffic of 630 E1s, carrying international traffic, in a period of 10 days.
- **Alternative B:** Applying focussed Speaker Identification, in combination with the VASTech Zebra Network Analysis capability, to search in a subset of the calls that will contain the target call more probably than the rest.

It is proposed that the Alternative B provides a much higher benefit:cost ratio than Alternative A.

In addition, it is found that Alternative A might very well be impractical in real conditions.

Accordingly it is proposed that a pilot system, based on Alternative B and as specified in this document on a high level, is accepted as compliance to the tender requirement.

In addition, it is suggested that this pilot system is used to further refine user requirements for later acquisition efforts, if required.

1 INTRODUCTION

1.1 Scope

This document:

1. Provides background information on Speaker Identification (SI) and provides parameters and examples that can be used to determine how practical it is to use Speaker Identification in a large passive surveillance system
2. Suggests a solution that might provide a better investment and provide practical results, depending on the criteria used by the Customer.
3. Provides a high level specification for a trial system that can be used to evaluate the suggested solution.
4. Provides, as appendices, additional source and background documents from leading SI vendors

This document also serves as input to a similar study on Language Identification.

2 GENERAL BACKGROUND ON SPEAKER IDENTIFICATION

2.1 High level definition

For this case, Speaker Identification is described as when a hardware and software is used to attempt to identify a target speaker in a set of calls that contain unknown speakers.

This is done by comparing the previously enrolled feature file of the target speaker with the features files of the unknown speakers in the all the calls. In the comparison process, a score is calculated to show how closely the known target speaker's characteristics correlate with those of the unknown speaker.

Calls with a score above predefined thresholds are defined as the Above Threshold Group (ATG) and most probably contain the target speaker. However, to verify that one or more of the calls actually contain the target speaker, it is required that a human operator has to listen through the calls in the ATG.

2.2 Speaker Identification is a probabilistic process and errors are involved

2.2.1 Probabilities

Speaker Identification is a probabilistic process and it is therefore *not certain* that a specific target is in a call – the target *might be* in a call in the ATG. Being a process based on probabilities, we can expect certain errors in identifying a specific speaker in a call.

2.2.2 Types of errors

The following types of errors exist:

- Miss Error (or called False Rejection Error, abbreviated FRR), and
- False Alarm Error (or called False Acceptance Error, abbreviated FAR).

If you do a Speaker ID test, you'll get scores of how well the known target's characteristics match with the unknown speaker. If you now set a decision threshold, and decide that all scores above the threshold indicate a target, while all scores below the threshold do not indicate a target, then you'll experience the type of errors in the following table. (These types of errors are present in all SID systems.)

	Speaker ID system *thinks target is not in call*, at a given threshold setting. Calls in this column all have scores below the threshold.	Speaker ID system *thinks target is in call,* at a given threshold setting. Calls in this column are all in the ATG
Target speaker is actually in call	Miss error (False Rejection error- FRR)	Good – no error
Target speaker is actually not in call	Good – no error	False alarm (False acceptance error -FAR)

2.2.3 Detection Error Trade-off curve (DET curve)

If one now repeats a Speaker Identification test at different decision threshold settings, and one plots the FAR against the FRR, one gets a curve such as in Figure 1.

This curve is called a DET curve (Detection Error Trade-off curve). The values in this curve are based on practical results (see figure 1 in stbu-taslp-07.pdf). The values in this curve in the referenced document are optimistic, since it refers only to a single language speaker. However, let us assume that the values in this curve are also applicable to the Customer's situation.

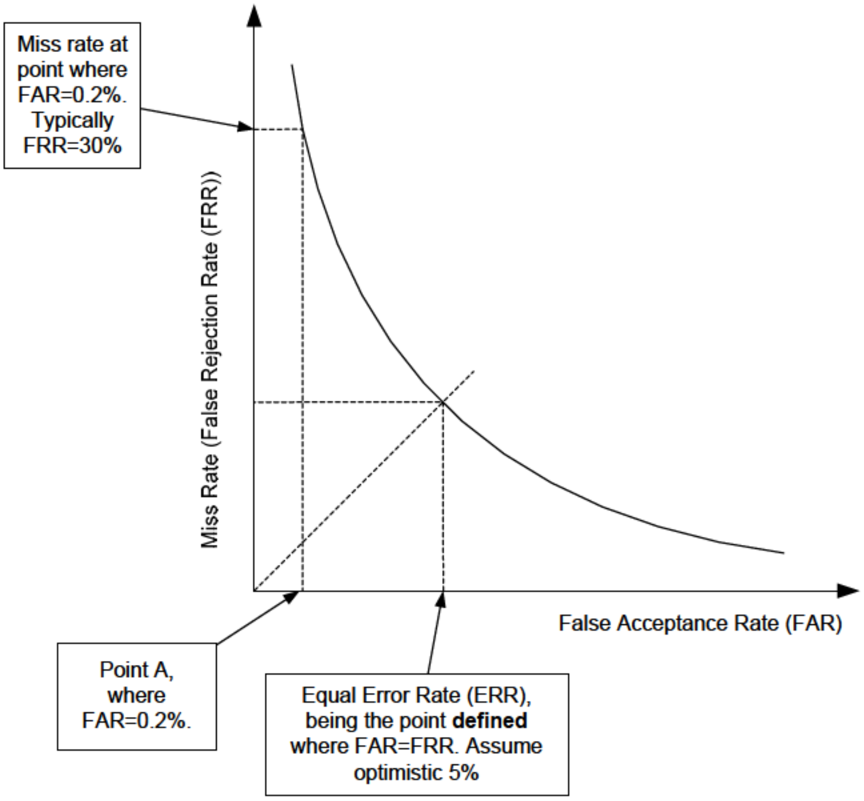

Miss rate at point where FAR=0.2%. Typically FRR=30%
Miss Rate (False Rejection Rate (FRR))
False Acceptance Rate (FAR)
Point A, where FAR=0.2%.
Equal Error Rate (ERR), being the point **defined** where FAR=FRR. Assume optimistic 5%

Figure 1: DET Curve - example

2.2.4 Equal Error Rate (EER)

The point on the curve where the FAR=FRR is defined as the Equal Error Rate (EER). This simply a matter of definition, and EER is used as one parameter to determine how good a system is, and to compare systems from different vendors. However, the EER must be carefully interpreted and it must be ensured that the EER's have been measured under exactly the same conditions between the different systems.

Consider now the EER and assume it is 5%. **Typically, for international calls, with short speech and where it is tried to compare the same speaker samples over different channels, the EER will be significantly worse – in the order of around 10-20% (see HP appendix).** For the purpose of this illustration however, assume the EER is 5%.

As illustration, assume there are 100 calls and only one of the calls contains target speaker X. If the threshold has been set at T_{EER}, one can expect that approximately 5% (since the FAR=5%) of the calls (i.e. 5 calls) will be above the threshold and hence will be returned in the Above Threshold Group (ATG). The calls in the ATG are deemed as the calls containing the target. This is clearly wrong since only one call contains the target. In addition, there is a 5% (FRR) risk that the target is not in the group of returned "target" calls.

The FRR can also be explained as follows:

Assume that the above test has been done 100 times on the same data with different speakers. After each test, one will get approximately 5% (i.e. 5) calls returned as groups of "targets" because the decision threshold has been selected where the FAR is 5%. The actual target may or may not be in the specific ATG. After running the test 100 times, one will have 100 ATGs of "targets", each group with the average size of 5 calls. Because of the FRR error, however, one will find that in 5% of these groups the actually target is not included and has been missed.

2.2.5 Discussing Speaker Identification in the case of 5'000'000 or more calls

Let us now move to a larger case with real values.

2.2.5.1 Assuming EER of 5% and FAR = FRR = 5% setting

Assume the decision threshold has been set at the EER point of 5%, i.e. 5% of the population of calls will be falsely accepted as targets, and 5% of the actual targets will be missed.

2.2.5.1.1 One target

Assume that we now search for one target. The target is in one call in a set of 5'000'000 calls (which is actually one day's of traffic). This means that the Above Threshold Group (ATG) will contain 5% (since the FAR=5% at this point) of 5 million calls, i.e. 250'000 calls. If one listens perfectly through the 250'000 calls, one has a 95% probability that the target will be found (since the FRR =5%, and hence there is a 5% chance that the target is not in the ATG).

(Obviously, if one is trying to find the specific target in a set of 50 million calls (being approximately 10 days traffic), one will have 2.5 million calls in the ATG, still with a probability of 5% that the target is not in this group.)

2.2.5.1.2 More than one target

Assume now that we have two targets and try to find them in the 5 million calls (each target has spoken, but only once). After doing the analysis for target 1, we'll have an Above Threshold Group (ATG) of 250'000 calls and after doing the analysis for target

2, we'll get another ATG comprising 250'000 calls. These two ATGs will definitely not be the same, and may, at worst actually consist out of 2 groups that contain completely separate calls. So, in the case of 2 targets, one will get at best a total of 250'000 calls (not likely) to up to 500'000 calls (worst case) to listen through. The exact total number of calls will be somewhere between these extremes since some calls may be in both groups at the same time.

If we now expand this challenge to where we have 10 targets, we'll get back 10 ATGs of 250'000 callers each. The total of calls to be listened through will be between the ranges 250'000 (just a theoretical limit – not practical) up to 2.5 million calls (also unlikely).

Clearly, in the case of 50 million calls and have 10 targets, the total answer groups will contain between 2.5 million calls up to 25 million calls (in the case where the answer groups contain completely separate calls). Even in this case, there is still a probability of 5% that a specific target (from the set of 10 targets) will not be in the specific target's answer group.

As we have more and more targets, the number of calls that one has to listen through (i.e. the sum of all the calls in all the ATG's) will approach the complete initial set of calls (ignoring now issues such as gender identification).

2.2.5.2 Change the threshold so that FAR is 0.2% to try to get more manageable results (less calls to listen through)

Consider now point A in Figure 1. At this point the threshold has been set such that the FAR is 0.2% and the FRR is approximately 35%. This is a trade-off, as the name of the DET curve implies, and is a fact of life: the lower the FAR, the higher the FRR.

Assume 50 million calls and one target. The answer set will contain a 100'000 calls, in which the target should be with a 65% probability (100% - 35% FRR).

Assume the case of 50 million calls, and 10 targets. The total answer sets can anything between 100'000 to 1 million calls. After listening through these calls, on average one will find that you've missed 3.5 targets, say 4 (being 10 targets x FRR rate of 35%).

2.2.5.3 Summarizing calls that one may have to listen through, for different FARs and different sizes of sets of calls

The following table summarizes a number of different scenarios. In the scenarios the following has been varied:

- Number of targets
- Number of calls in the call set (calls that may contain the target and that has to analyzed)
- FRR and FAR

Table 1: Results in the case of brute force Speaker Identification

A	B	C	D	E	F
Line	**Description**	**Scenario**			
1	FAR (False alarm probability) [%]	5	5	0.2	0.2
2	FRR (miss probability) [%]	5	5	35	35
3	Assumed time to listen, per call, to verify [seconds]	30	30	30	30
4	**Number of calls to be searched through**	**5,000,000**	**50,000,000**	**5,000,000**	**50,000,000**
5	**Number of targets to be found**	**1**	**1**	**1**	**1**
6	Number of calls in Above Threshold Group per target (ATG)	250,000	2,500,000	10,000	100,000
7	Time required to listen through complete ATG [hours]	2,083	20,833	83	833
8	Probability of finding target in ATG	95	95	65	65
9	Probability that target is missed	5	5	35	35
10	**Number of targets to be found**	**10**	**10**	**10**	**10**
11	Number of calls in Above Threshold Group per target (ATG)	250,000	2,500,000	10,000	100,000
12	Maximum number of calls in of all ATGs, together	2,500,000	25,000,000	100,000	1,000,000
13	Time required to listen to max size of ATG [hours]	20,833	208,333	833	8,333
14	Mandays/shifts required to listen through max size ATG [days] at 6 hours continuous per day/shift	3,472	34,722	139	1,389
15	Expected number of targets found	9.5	9.5	6.5	6.5
16	Targets missed, after listening through all	0.5	0.5	3.5	3.5
17	**Number of targets to be found**	**100**	**100**	**100**	**100**
18	Number of calls in Above Threshold Group per target (ATG)	250,000	2,500,000	10,000	100,000
19	Maximum number of calls in of all ATGs, together	5,000,000	50,000,000	1,000,000	10,000,000
20	Time required to listen to max size of ATG [hours]	41,667	416,667	8,333	83,333
21	Mandays/shifts required to listen through max size ATG [days] at 6 hours continuous per day/shift	6,944	69,444	1,389	13,889
22	Expected number of targets found	95	95	65	65
23	Targets missed, after listening through all	5	5	35	35

The following conclusions can be made from the above table:

- If one attempts to find more targets in a specific call set, then the work to find the target drastically increases
- If one tries to reduce the number of calls one has to listen through, by changing the threshold to reduce the FAR, then one significantly increases the risk in missing the target in any case. (See cell E18-E23: if you listen through all the calls, you can expect to find 65 out of the 100 targets and still miss 35 targets). Working at a FAR =0.2%, the set of calls that you have to listen through to try to find the 100 targets could be as large as 1 million calls, if one started with a possible 5 million calls in which the target might be. Similarly, if you start with 50 million calls and try to find the 100 targets, one might have to listen through up to 10 million calls and still miss 35 of the targets, on average.

It should be noted that the above table is based on a system with an EER of 5%. Practically, in the case of international calls, the results may be worse. See HP document attached, referring to EER of between 10 and 20%.

2.3 Suggested conclusion on practicality of Speaker Identification in massive system

The above analysis and tables are sufficiently accurate to suggest conclusions – the fact that some of the ATGs may overlap is offset by the fact that the actual EER may be up to 4 times worse.

One can conclude that:

- One has to spend vast capital resources (hardware and software) and vast running costs (manpower to listen through a large number of calls) to try to find targets in a large set of calls if one only applies automatic Speaker Identification.
- Even after the investment and expenditure, one may still miss a significant number of targets

Some may argue that it is prohibitively expensive, while offering low value, to apply only "raw" speaker identification on a massive number of calls.

This is due to the fact that Speaker Identification is probabilistic, and one is trying to find a single needle in a very large haystack.

2.4 Recommendation

It is recommended that Speaker Identification is **not implemented on its own on the total number of calls**

A possible approach will be discussed in the next section, which may be more economical and provide better value for money.

3 ANOTHER SOLUTION, POSSIBLY MORE ECONOMICAL AND PRACTICAL, DEPENDING ON REQUIREMENTS

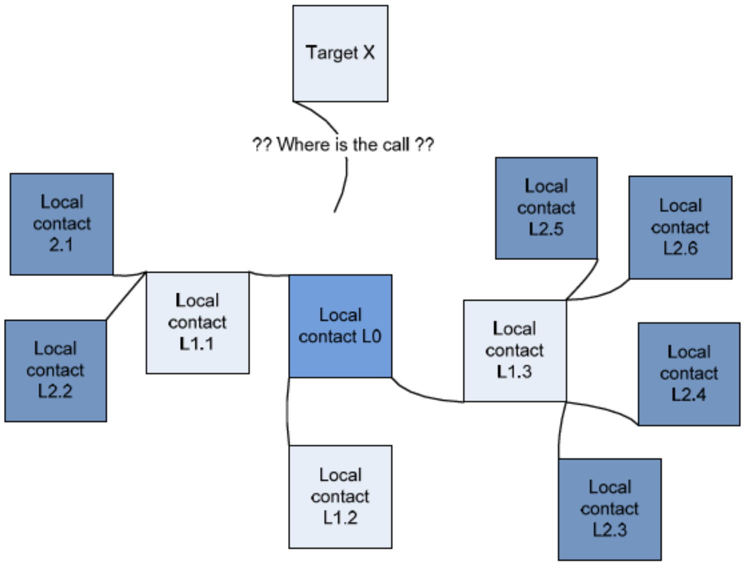
3.1 Task and assumptions

Assume a target X is in foreign country A. The task is to find the calls that contain target X. The telephone number of target X is unknown, due to various possible reasons:

- The calls of X could be routed through international switches where his number has been stripped off
- Target X might actually be visiting country B, or C and make a call from any of these countries.

The following is further assumed:

- A sufficient sample of speech of Target X is available to make a feature file (voice print) of X.
- Target X has spoken already to a telephone number in the customer's country and this local number is known, say this is L0.
- It can be expected Target X will very likely speak again with local number L0, or with the contacts of L0 (say these are called L1), or with the contacts of L1 (say these are called L2). In other words, it can be expected that Target X will again contact L0 or his contacts, or their contacts, etc.



Assume that the Zebra Network view is used, with three layers as indicated in a very simplified manner. It can be expected that X will speak to one of the contact in this network model again.

The following is a rough assumption, for illustrative purposes, of the total number of calls that occur between the layers L0-L1-L2. It can clearly be different in real life, but it should also be clear that the number of calls in which to search Target X is drastically reduced.

Number of links from L0-L1, relevant during the period under investigation	50
For each L1, the average number of links between L1 -L2, relevant during the period	50
Number of local parties to be monitored (L0-L1-L2)	2,501
Average number of calls related to each monitored party during period (note some are duplicated)	100
Total number of calls to be investigated	250,100

Using the above number of calls, the following can be determined, given one try to find one target.

Table 2: Results in the case of assisted Speaker Identification

A	B	C	E
Line	Description	Scenario	
1	FAR (False alarm probability) [%]	5	0.2
2	FRR (miss probability) [%]	5	35
3	Assumed time to listen, per call, to verify [seconds]	30	30
4	**Number of calls to be searched through**	**250,100**	**250,100**
5	**Number of targets to be found**	**1**	**1**
6	Number of calls in Above Threshold Group per target (ATG)	12,505	500
7	Time required to listen through complete ATG [hours]	104	4
8	Probability of finding target in ATG	95	65
9	Probability that target is missed	5	35

Comparing the above table with Table 1, one finds a big reduction of calls to be listened through. For example: at FAR=5%, the calls to be listened through is decreased from 2.5 million or 250'000 (depending on the period of interest) down to 12'500 calls.

3.2 Conclusion

Depending on the validity of the assumptions above, alternative approaches exists to assist the customer more cost-effectively and practically to find a target than simply applying Speaker Identification on all 630 E1s.

One specifically attractive alternative is to use additional intelligence, such as gathered through the Zebra network analysis to find the numbers that could possibly be called by Target X. This provides a much more focussed search, with the search focussed in the most likely area, as opposed to using brute force over all calls.

Other alternatives may also exist.

4 SUGGESTED PILOT SYSTEM

4.1 High level specification

4.1.1 Included functionality

Provide a pilot system, that enables the following work flow:

- For Target X:
  - Generate a feature file (voice print) for a specific target X.
  - Determine a filter that can filter out the calls that most probably will contain target X. This can be done by:
    - Using such as network analysis, to determine the parties that a specific target may call.
    - Using any other available intelligence.
  - Export all calls related to the specific target filter (say this is called the target set) to the Speaker ID system. It must be able to export compressed files, which shall automatically be decompressed prior to analysis. It is accepted that Speaker Identification on decompressed files may lead to unreliable/low quality results.
- Similarly, create filters for up to 20 targets and export the calls related to each target filter. The total number of calls to be exported shall not exceed more than the equivalent traffic of 10 stereo E1, i.e. 20 Mbit/s.
- Analyse each target set and rank the calls in each target set with the probability that it contains the specific target
- Listen through the calls in each target set (by clicking on the call link), from the calls with the highest score, to attempt to find the target. Typically, in a system with and ERR of 10% (due to e.g. the call quality and length of calls), one could expect to find the target with 90% probability in the top 10% of the scored calls.

4.1.2 Excluded or limitations in the case of the pilot system

- Feature files (voice prints) will not be stored
- Only the set of Languages standard to the installed software shall be supported

4.2 Acceptance tests

- Acceptance test, due to the fact that the exact ERR is not known for the specific call circumstances: (as described in the ATP).
- Acceptance tests shall not be conducted on Language Identification in the case of the pilot system.
- Due to development required, the acceptance test should only be executed at Provisional Acceptance.

4.3 Benefits of Pilot System

- Allows the development of detailed user requirements for future implementation and expansion
- Allows direct assessment by the customer of the usability and benefit of Speaker Identification.

--end--