Two-Wire Nuisance Attribute Projection

Yosef A. Solewicz1, Hagai Aronowitz2

1 Department of Computer Science, Bar-Ilan University, Ramat-Gan, Israel
2 IBM Haifa Research Labs, Haifa 31905, Israel

solewicz@013.net, hagaia@il.ibm.com

Abstract

This paper addresses the task of nuisance reduction in two-wire speaker recognition applications. Besides channel mismatch, two-wire conversations are contaminated by extraneous speakers which represent an additional source of noise in the supervector domain. It is shown that two-wire nuisance manifests itself as undesirable directions in the inter-speaker subspace. For this purpose, we derive two alternative Nuisance Attribute Projection (NAP) formulations tailored for two-wire sessions. The first formulation generalizes the NAP framework based on a model of two-wire conversations. The second formulation explicitly models the four- vs. two-wire supervector variability. Preliminary experiments show that two-wire NAP significantly outperforms regular NAP in varied two-wire tasks.

Index Terms: speaker recognition, NAP, two-wire.

1. Introduction

The task of speaker recognition in two-wire test sessions can be found in a variety of scenarios, such as meetings, call-centers or network monitoring. The typical approach in these cases is applying some kind of speaker segmentation as a pre-processing stage, before applying speaker recognition [1]. Lately, [2] proposed to skip the segmentation stage, directly training and/or testing on two-wire session. However, the accuracy attained through this simplistic approach was not satisfactory. In a previous paper [3], we proposed several techniques for enhancing two-wire recognition without explicitly using speaker segmentation. In this paper, we derive two alternative supervector subspace removal formulations which attempt to improve a direct two-wire speaker recognition task. In particular, we initially derive a generalization of the well-known NAP technique [4] taking into account also inter-speaker variability factors, in addition to intra-speaker variability effects, as currently approached in NAP. We show that in the supervector domain, two-wire noise which is caused by an interfering speaker (the conversation partner) is rooted on the speaker space which should be therefore partially removed. Alternatively, we show that it is possible to explicitly model four- vs. two-wire supervector variability within the traditional NAP framework. The paper is organized as follows. Section 2 presents the theoretical foundations of the proposed methods. In Section 3, we report validation experiments used to corroborate our assumptions. In Section 4, we conclude the paper and offer future research directions.

2. Two-wire NAP approaches

In this session we present two NAP approaches particularly designed to encompass two-wire variability, in addition to standard channel nuisance. Specifically, our goal is to estimate and remove supervector directions containing two-wire nuisance, i.e., those directions mostly affected by the undesired effect of jointly MAP-adapting an extraneous speaker together with the target one. Two approaches are suggested to estimate the two-wire nuisance subspace. Initially, we address this issue by proposing a mathematical model for two-wire conversations and then re-deriving the NAP formulation for this model. Alternatively, we suggest a modification of the NAP framework, using an explicit estimation of four- vs. two-wire (instead of the typical inter-session) variability.

2.1. Mathematical variability model

The original NAP technique [4] aims at deriving a nuisance subspace, typically reflecting intra-speaker variability, which is later projected-out from each conversation. Our objective is to derive a two-wire NAP formulation, which considers also the “noise” introduced by undesirable speakers in two-wire conversations.

State-of-the-art speaker recognition systems model a conversation as a “supervector”, which is in fact a concatenation of the UBM MAP-adapted GMM means obtained processing this conversation [5]. Suppose $X_i$ and $X_j$ are the supervectors estimated over four-wire conversations of speakers $i$ and $j$, respectively. Now, suppose a hypothetical two-wire conversation between speakers $i$ and $j$. We would like to express the supervector obtained for this two-wire conversation in terms of both four-wire supervectors $X_i$ and $X_j$. A reasonable approximation is to assume that MAP adaptation is a linear operation and therefore the resulting two-wire supervector could be modeled by a linear combination between $X_i$ and $X_j$. Thus, $X'_i$, the two-wire supervector version of $X_i$, contaminated by the interfering supervector $X_j$ is given by:

$$X'_i = w X_i + (1-w) X_j$$

(1)

where $w \in (0,1)$ defines the dominance ratio of speaker $i$ in the two-wire conversation. For simplicity, we consider $w = 0.5$ for all speakers and drop it henceforth.

Within the NAP formulation, the supervector differences between the sessions of a given speaker and the mean across these sessions are pooled across several speakers forming an intra-speaker variability matrix $M$. The principal eigenvectors of the covariance of $M$ span the intra-speaker variability subspace, which should be subsequently eliminated from the conversation supervectors. Formally,

$$M = [X_{1,1} - \bar{X}_1 \cdots X_{1,1} - \bar{X}_1 \cdots X_{1,1} - \bar{X}_1]$$

(2)
where $X_{i,1\ldots Ni}$ are supervectors derived from speaker $i$, sessions 1 to $Ni$, and $\bar{X}_i$ represents the mean supervector of speaker $i$, given these sessions.

In order to derive the two-wire NAP formulation, we replace each four-wire supervector $X_i$ in Eq. (2) by $\bar{X}_i$, as defined in Eq. (1). After some simplification, we obtain a matrix $M'$ which models the intra-speaker variability space of two-wire conversations.

$$M' = [\cdots X_{i,1\ldots Ni} - \bar{X}_i + X_{i,j\neq i,1\ldots Ni} - \bar{X}_{j\neq i}\cdots]$$

where $\bar{X}_{j\neq i}$ stands for the mean supervector among the sample speaker population, excluding speaker $i$.

The first difference term of $M'$ reflects the original intra-speaker variability as for $M$ and for simplicity we denote it by $A$. In contrast, the second difference term accounts for inter-speaker variability factors, which we represent by $B$. Thus,

$$M' = [\cdots A_i\cdots] + [\cdots B_{j\neq i}\cdots]$$

Finally, to derive the two-conversation NAP formulation, we must compute the eigenvectors of the covariance of $M'$. Unfortunately, this calculation is not an easy task. It involves at first explicitly summing the covariance matrices of the cross-terms of $A$ and $B$, which are extremely high-dimensional (square of the supervector length). Nevertheless, if we assume that the intra- and inter-speaker terms are statistically uncorrelated, $M'(M')^T$, the covariance matrix of $M'$ can be approximated as:

$$M'(M')^T \approx AA^T + BB^T \approx (A + B)(A + B)^T$$

In this case, we can obtain the desired eigenvectors as a function of the eigenvectors of $(A + B)/(A + B)$ which is a square matrix of much smaller dimension and avoids an awkward straightforward calculation of the covariance of $M'$ as defined in Eq. (4).

(In fact, the assumption that the intra- and inter-speaker subspaces are orthogonal is not true in practice. It will be commented later that the principal eigenvectors of both subspaces possess a reasonable degree of correlation.)

In summary, two-wire NAP involves projecting-out not only the principal intra-speaker directions (or the eigenchannels spanning the “channel” space), as in traditional NAP, but rather the principal directions of the sum of the intra- and inter-speaker space. At first glance, removing the inter-speaker (or simply, the “speaker”-) space seems to be counter-intuitive. This phenomenon may be justified as follows. Regular eigenspace decomposition aims at finding the principal directions spanning the desired data, so that the higher-order components of the new coordinate axes would mainly contain noise. This is also the case in speaker recognition, where it was found to be advantageous under certain circumstances to preserve only the principal components of the speaker-space, known as “eigenvoices” [6]. However, in two-wire conversations, the partner speaker turns out to be a major source of noise in the supervector space. Moreover, this source of noise, being an outsider speaker lies exactly in the speaker space. Thus, projecting-out some of the speaker space from two-wire conversations should in theory alleviate this effect. On the other hand, excessively removing the speaker space might harm the inter-speaker discrimination functionality and eventually, a compromise must be achieved.

### 2.2. Explicit modeling of variability

In the previous session, we have developed a mathematical model for two-wire supervectors as a weighted sum of two four-wire supervectors and further derived its corresponding NAP formulation. In this session, we propose to directly build a NAP matrix modeling the four- vs. two-wire variability, instead of the usual inter-session variability.

Specifically, we form the variability matrix $M$ as in Eq. (2) with the differences between four- and two-wire supervectors of a same speaker. In particular, given a two-wire conversation, we calculate its two-wire supervector and those corresponding to both four-wire sides. Then, we obtain two supervectors expressing the differences between the two-wire and each of the four-wire supervectors. The variability matrix is subsequently populated with various pairs of different supervectors, as noted in Eq. (6), where $X_{2Wi}$ stands for a two-wire supervector of speaker $i$ and $X_{4Wa,i}$ and $X_{4Wb,i}$ are the respective four-wire sides of the same conversation. Afterwards, the eigenvectors of the covariance matrix of $M$ are estimated. They should hopefully span the four- vs. two-wire variability sub-space to be later projected-out from the supervectors.

$$M = [\cdots X_{2Wi,j} - X_{4Wa,i}, X_{2Wi,j} - X_{4Wb,i}\cdots]$$

To a certain extent, matrix $M$ in Eq. (6) resembles the term $B$ in Eq. (4), fairly reflecting inter-speaker variability. To be precise, while $B$ models the standard inter-speaker variability (i.e. the dissimilarity between a speaker and a virtual “average” speaker), matrix $M$ here models the variability between a speaker and a mixture of speakers, including himself.

Unlike the previously presented mathematical approach, in which inter-session variability (term $A$ in Eq. (4)) is a natural outcome of the formulation, it is not being automatically addressed in the current approach, since the four- vs. two-wire variability is estimated over a unique session for each speaker. Therefore, we can further expand the nuisance sub-space defined by $M$, adding the regular inter-session variability nuisance component, as defined by $M'$ in Eq. (7).

$$M' = [\cdots A_i\cdots] + M$$

Finally, we can estimate the eigenvectors of the covariance of $M'$, delineating the general nuisance sub-space, just as described earlier for the mathematical modeling approach, in Eq. (5).

### 3. Experimental validation

#### 3.1. System setup and protocols

The speaker recognition system used in all the reported experiments is a GMM-NAP-SVM supervector system quite similar to the one described in [7]. The GMM subsystem is based on [5] and the number of Gaussians is 512. Each supervector is normalized by the corresponding standard deviation of the UBM Gaussians. We use LIBSVM [8] to implement the SVM subsystem, which renders the actual recognition score, with no further normalizations applied. The front end is based on Mel-frequency cepstrum coefficients.
(MFCC). An energy-based voice activity detector is used to locate and remove non-speech frames. The final feature set consists of 13 cepstral coefficients augmented by 13 delta cepstral coefficients extracted every 10 ms using a 25 ms window. Finally, feature warping [9] is applied with a 300 sample frame window.

The experiments reported in this paper consist of the male subset of the NIST-2005 SRE protocol [10]. There are at all 274 target speakers and 965 conversation tests, producing 951 target and around 8000 impostor scores. Moreover, since this evaluation is designed for four-wire sessions, in order to obtain two-wire testing sessions for our experiments, we artificially sum the two sides of the testing conversations in the original protocol. (In fact, the speaker present in the summed side might be eventually the same speaker present in the specific trial model. Therefore, in these cases, original four-wire impostor trials mistakenly turn out to be two-wire target trials. Since the evaluation key files lack this piece of information, we dropped a few trials for which the corresponding additional side attains higher recognition score than the maximum score obtained using the original four-wire impostor trials.)

The NIST-2004 and 2006 SREs corpora [10] were used as development datasets. Specifically, SRE’04 was used for training the UBM (240 conversations), SVM background modeling (200 conversations) and for inter-session variability modeling (124 male speakers, 1457 conversations in total), while the SRE’06 dataset was used for inter-speaker space modeling (350 conversations).

3.2. Speaker- and channel-space orthogonalization

We have shown through two different approaches that two-wire conversations possess an inter-speaker nuisance term, in addition to the traditional intra-speaker variability component. In order to estimate the general nuisance subspace, we assumed in Sections 2.1 and 2.2 that both components are statistically uncorrelated. As a matter of fact, a direct implementation of the proposed methods through Eq. (5) attained poor results. A subsequent analysis showed that the eigenvector sets of the channel and speaker space possess a reasonable correlation degree which probably invalidates the assumption that channel and speaker space are uncorrelated, which is used to derive Eq. (5). We therefore propose an alternative formulation in order to estimate the combined channel and speaker space eigenvectors. This is performed by pooling the respective eigenvectors of the covariance matrices of $A$ (channel space) and $B$ (speaker space) in Eq. (4) and subsequently orthogonalizing the combined eigenvector space. In other words, we approximate the general two-wire nuisance space by the space jointly spanning the low-rank channel and speaker subspaces. Formally, let $\{a_1,\ldots,a_m\}$ be the first $m$ eigenvectors of $AA'$, and $\{b_1,\ldots,b_n\}$ be first $n$ eigenvectors of $BB'$, which span the channel and speaker space respectively. We append both bases, forming the combined subspace basis $\{a_1,\ldots,a_m,b_1,\ldots,b_n\}$ and apply the Gram-Schmidt process to this vector set, obtaining $\{a_1,\ldots,a_m,\tilde{b}_1,\ldots,\tilde{b}_n\}$.

The channel space component, $\{a_1,\ldots,a_m\}$ remains unaltered, since it represents an already orthonormal basis, however $\{\tilde{b}_1,\ldots,\tilde{b}_n\}$ span the subspace of the speaker space which is orthogonal to the channel space.

This new set of eigenvectors actually constitutes the combined intra- and inter-speaker variability NAP matrix to be used in two-wire tasks. Note that this novel NAP matrix is identical to the basic NAP up to the channel eigenspace rank ($m$), but contains $n$ extra eigenvectors, which we call the orthogonalized eigenvoices (OEVs).

Specifically, we estimate the regular NAP (channel) eigenspace using NIST SRE’04 and the speaker eigenspace using NIST SRE’06. We have observed that $m$ should be relatively large in order to better estimate the orthogonalized speaker space. We therefore append the eigenvoices ($n=300$) to the eigenchannels ($m=400$) and obtain novel orthogonormalized eigenvectors through the Gram-Schmidt process.

The same procedure described above is used to obtain the combined nuisance space defined by the explicit method presented in Section 2.2. In particular, the eigenvectors of the covariance of $M$ in Eq. (7) are appended to the eigenchannels and jointly orthogonalized as explained. For simplicity, we also refer to these eigenvectors, as OEVs. (In fact, we have already commented on the similarity between the eigenvectors originated from matrix $M$ in Eq. (7) and the traditional eigenvoices.)

3.3. Results

In this section, we report experiments performed in order to evaluate and compare the proposed theoretical and explicit two-wire NAP approaches in distinct tasks involving four-wire, two-wire and segmented conversations.

Specifically, in these experiments, the NAP used contained 400 eigenvoices followed by 300 OEVs, as explained in Section 3.2. For simplicity, we always apply the whole channel space projection to training supervectors. Subsequently, we progressively increase the OEV rank which is also projected-out from these supervectors.

Initially, as expected, we observed that speaker space projection has no effects on the full four-wire task. We then evaluated the two-wire NAP methods on four-wire training vs. two-wire testing task. Figure 1 shows the detection cost function DCF [10] performance evolution obtained by gradually increasing the OEV rank appended to the eigenchannel set for both the mathematical and explicit approaches. For comparison, the results of merely increasing the eigenchannel rank (NAP) is also showed.

![Fig. 1. DCF for distinct OEV ranks, for the 4w-2w task](image-url)
It can be clearly seen that further increasing channel space projection is not effective, while speaker space projection does improve recognition performance.

The next experiment analyzes the four-wire training vs. speaker-segmented testing. Typically, speaker-segmented conversations are derived from the respective summed-side conversations by means of some segmentation system. Two scores are obtained, one for each of the segmented excerpts. The overall detection score is the maximum between both scores. Actually, the maximum function itself, which is needed when simultaneously testing two hypothesized speakers, introduces an intrinsic detection performance drop, as commented in [11]. Note that this increase in error happens even with perfect segmentation (i.e. using both four-wire sides) which represents the upper bound baseline for this task. (The two-wire testing task, discussed earlier essentially represents a lower bound for performance.) Figure 2 depicts performance evolution using perfect segmentation. Interestingly, OEV projection proved to be effective even though the testing conversations are actually perfectly segmented.

As a control experiment, we used a standard segmentation system in testing. Similar trends were observed but used of distinct segmentation systems are required before further conclusions. Generally, it can be observed that less OEVs are required in this task as compared to the previous two-wire testing conversations, which are heavily contaminated by a second speaker.

We finally analyze the full two-wire condition. The results are shown in Figure 3. In this task, the explicit two-wire NAP approach clearly outperforms the mathematical one. Moreover, as could be expected, the advantages of two-wire NAP approach clearly outperforms the mathematical approaches proposed. One based on a mathematical eigenspaces, to form the ultimate nuisance space. Two orthogonalize the pooled channel and speaker variability factors. The optimal implementation found so far is to orthogonalize the channel space projection does improve recognition performance.

5. Discussion and conclusions

In this paper we developed and evaluated novel formulations for the NAP framework, which address two-wire or segmented conversation tasks. We showed that such formulations include, besides the regular intra-speaker variability term, an additional term representing inter-speaker variability factors. The optimal implementation found so far is to orthogonalize the pooled channel and speaker eigenspaces, to form the ultimate nuisance space. Two approaches were proposed. One based on a mathematical model of two-wire conversations and the other based on explicit two-wire variability modeling. Distinct recognition tasks were evaluated and the latter approach clearly outperformed the former one. Inspired in the strength of explicitly modeling four-vs. two-wire supervector variability, we have successfully evaluated the task of automatic classification between these two supervector classes. This may lead to the development of methods able to automatically assess the segmentation quality of two-wire conversations. Additional future work should focus on the combination of the proposed two-wire NAP with speaker segmentation techniques. Moreover, further research regarding the aspects of the speaker and channel space interaction should contribute to enhance the proposed framework.

![Fig. 2. DCF for distinct OEV ranks, for the 4w-seg. task](image)

Fig. 2. DCF for distinct OEV ranks, for the 4w-seg. task

![Fig. 3. DCF for distinct OEV ranks, for the 2w-2w task](image)

Fig. 3. DCF for distinct OEV ranks, for the 2w-2w task

5. References