AB/BA analysis: A framework for estimating keyword spotting recall improvement while maintaining audio privacy

Raphael Petegrosso, Vasistakrishna Baderdinni, Thibaud Senechal[*]**, Benjamin L. Bullough**[*]
Amazon / USA
{petegrr,vbaderdi,thibauds,bullough}@amazon.com

Abstract

Evaluation of keyword spotting (KWS) systems that detect keywords in speech is a challenging task under realistic privacy constraints. The KWS is designed to only collect data when the keyword is present, limiting the availability of hard samples that may contain false negatives, and preventing direct estimation of model recall from production data. Alternatively, complementary data collected from other sources may not be fully representative of the real application. In this work, we propose an evaluation technique which we call AB/BA analysis. Our framework evaluates a candidate KWS model B against a baseline model A, using cross-dataset offline decoding for relative recall estimation, without requiring negative examples. Moreover, we propose a formulation with assumptions that allow estimation of relative false positive rate between models with low variance even when the number of false positives is small. Finally, we propose to leverage machine-generated soft labels, in a technique we call Semi-Supervised AB/BA analysis, that improves the analysis time, privacy, and cost. Experiments with both simulation and real data show that AB/BA analysis is successful at measuring recall improvement in conjunction with the trade-off in relative false positive rate.

1 Introduction

Keyword spotting (KWS) is the task of identifying if one of a set of keywords, also called wakewords, is present in a speech segment. It is the gatekeeper component that enables hands-free interaction with many smart assistants using voice-enabled smart devices, such as Amazon Echo, Google Home, and Apple HomePod. Extensive work has been done to develop and improve KWS performance, including improvements in architecture (Chen et al., 2014; Sun et al., 2017; Shan et al., 2018; Wu et al., 2018;

---
[*]Corresponding authors

Gao et al., 2020), training efficiency (Tucker et al., 2016; Raju et al., 2018), as well as audio front-end (AFE) algorithms (Chhetri et al., 2018).

With many lines of research aiming to improve the quality of the KWS, there is also growing interest in techniques to measure if such new technologies are able to improve the customer experience, but less research attention has been given to this evaluation topic. One challenge is that KWS systems are designed to maximize user privacy by only collecting data when the keyword is identified. As a result, evaluations done using this biased dataset do not allow direct measurement of gains in recall metrics to determine if a new model is better than a baseline. Alternatively, evaluations can also make use of datasets not collected by the KWS, such as media recordings, background noise, and environmental sounds. However, these datasets may not be fully representative of the user experience from the evaluation point of view.

Prior research, such as from (Gao et al., 2020; Sainath and Parada, 2015), has made use of datasets with positive and negative labels in order to evaluate, respectively, gain invariant KWS models and CNN for small-footprint KWS. However, the negative data essentially consists of either negative labels obtained from data accepted by the previous models, or datasets composed of just background noise and noise from environment. In a different application context, (Miller et al., 2018) estimates the recall and the derivative of the precision with respect to the recall by modeling the unseen data distribution according to underlying assumptions. This distribution, however, is not clearly defined in the context of KWS when a variety of sounds, noisy conditions, and reverberation can occur.

In this paper, we present a new evaluation framework called AB/BA analysis. To the best of our knowledge, our work is the first to explore the problem of estimating recall improvement when only accepted data is available, such as in a KWS system

with rigorous privacy settings, in conjunction with the trade-off with false positive rate, without underlying assumptions on the data distribution. We show that the AB/BA analysis is a cross-model offline evaluation framework. In this framework, data is collected from two KWS models, say a baseline, A, and a candidate, B. By running offline model A through data collected by model B, and model B through data collected by model A, we are able to calculate relative metrics without the need for data not seen by individual models. We also present assumptions that can be applied to the relative metrics calculation that result in a lower variance estimator, even if the number of false positive is small. We additionally describe a Semi-Supervised formulation of AB/BA analysis which provides improvements in the analysis time, cost, and data privacy by utilizing soft machine-generated labels instead of human annotations from the models being evaluated. We present experiments using simulation data, real data, and also experiments comparing the performance of AB/BA and Semi-Supervised AB/BA.

2 Methods

In this section, we describe the proposed method for estimation of recall improvement. Basic metrics concepts can be found in Appendix A.1.

2.1 AB/BA Analysis

The AB/BA analysis framework is composed of four main steps, as depicted in Figure 1: Online data collection, Offline decoding, Labeling, and Metrics computation.

In order to collect data, given two KWS models, say A and B, the models are deployed simultaneously to two populations of users, also called A and B, as shown in Figure 1 (A). The percentage of users in each model is usually based on the presumed risk of deploying each model, as well as the statistical significance desired for the metrics computed, as shown later. It is important, though, that models are deployed simultaneously, with random assignment, similar to a conventional A/B-Test. Notice that the data collected by the models will only contain samples where the keyword is detected in order to preserve user privacy.

The collected data is then used for offline decoding, in which models are run offline on the data collected online. As shown in Figure 1 (B), the data collected by model A is used by model B for

keyword spotting, and model B is run on the data collected by the model A. A detailed review on keyword spotting can be found in (López-Espejo et al., 2021). Running a keyword spotter on an utterance i will produce a score s_i representing how likely it is to have detected the keyword.

The data that is collected and decoded offline by the KWS systems will also require labeling in order to be used for metrics estimation. This step, shown in Figure 1 (C), is usually done by either human annotation, or machine-generated. The process to use machine-generated labels is described in more detail in Section 2.2.

Human annotation is an expensive and time consuming process, in addition to being susceptible to error. Therefore, it is desirable to carefully select which utterances to annotate. Stratified sampling can be used to provide more annotations on models disagreement to reduce the need for human annotations (details in Appendix A.2). The labeled data is then used to compute metrics that represent the relative improvement of model B with respect of a baseline model A, as shown in Figure 1 (D).

AB/BA analysis utilizes two relative metrics: False Positive Rate Ratio and Recall Ratio. The Recall Ratio (rRecall) is a relative metric used for comparing two models to determine which model yields better recall. Given two KWS models A and B with datasets containing utterances collected by the same models online, and labeled as $L \in \{0, 1\}$, using the Bayes' theorem, the rRecall can be defined as:

$$
\begin{aligned}
rRecall &= \frac{P(B = 1 | L = 1)}{P(A = 1 | L = 1)} \\
&= \frac{P(B = 1 | A = 1, L = 1)}{P(A = 1 | B = 1, L = 1)},
\end{aligned} \quad (1)
$$

where $P(B = 1 | A = 1, L = 1)$ indicates the probability model B found the keyword when run offline ($B = 1$) on true positives from model A used online ($A = 1, L = 1$).

Therefore, a key aspect of AB/BA analysis is that the rRecall can be computed using terms $P(B = 1 | A = 1, L = 1)$ and $P(B = 1 | A = 1, L = 0)$ that are directly observable.

Concretely, we can compute rRecall using the following quantities:

$$
rRecall = \frac{NTP_{BA\_on\_A}}{NPos_A} * \frac{NPos_B}{NTP_{AB\_on\_B}}, \quad (2)
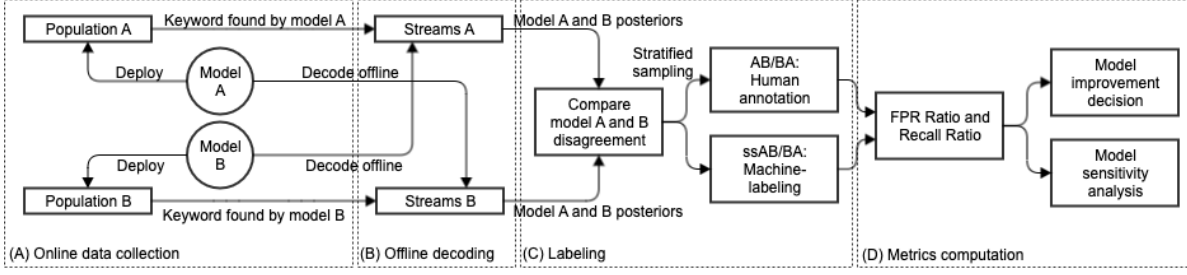$$

Figure 1: **The four main components of the AB/BA analysis**

where $NTP_{AB\_on\_B}$ is the number of TPs of model A and B on data collected by model B and $NPos_B$ is the number of positive labels on the data collected by B.

Similarly, the FPR Ratio (rFPR) between two models A and B can be calculated as:

$$rFPR = \frac{NFP_{BA\_on\_A}}{NNeg_A} * \frac{NNeg_B}{NFP_{AB\_on\_B}}, \quad (3)$$

where $NFP_{AB\_on\_B}$ is the number of FPs of model A and B on data collected by model B and $NNeg_B$ is the number of negative labels on the data collected by B.

Notice that the number of FPs can be small, leading to large variance in the rFPR estimation. Therefore, assuming that keywords and confusing sounds (those that induce FPs) generated by population A and population B are randomly drawn from the same distribution, we propose to assume that the ratio of TPs and FPs in the streams accepted by both models is the same, represented as:

$$\frac{NFP_{BA\_on\_A}}{NTP_{BA\_on\_A}} \approx \frac{NFP_{AB\_on\_B}}{NTP_{AB\_on\_B}} \quad (4)$$

Then, by introducing the following variables:

$$
\begin{aligned}
NTP_{AB} &= NTP_{BA\_on\_A} + NTP_{AB\_on\_B} \\
NFP_{AB} &= NFP_{BA\_on\_A} + NFP_{AB\_on\_B} \\
\alpha &= \frac{NTP_{BA\_on\_A} + NFP_{BA\_on\_A}}{NTP_{AB} + NFP_{AB}} \quad (5) \\
\beta &= \frac{NTP_{AB\_on\_B} + NFP_{AB\_on\_B}}{NTP_{AB} + NFP_{AB}}
\end{aligned}
$$

We can find rFPR and rRecall using:

$$
\begin{aligned}
rFPR &= \frac{\alpha(NFP_B + \beta NFP_{AB})}{\beta(NFP_A + \alpha NFP_{AB})} \\
rRecall &= \frac{\alpha(Nmiss_A + \beta NTP_{AB})}{\beta(Nmiss_B + \alpha NTP_{AB})},
\end{aligned}
\quad (6)
$$

where $Nmiss_A$ is the number of TPs accepted by model B but not accepted by A. With this estimator, uncertainties on NFP_{AB} have less impact on the rFPR uncertainty. This is shown in the simulation on Section 3.1.1.

2.2 Semi-Supervised AB/BA Analysis

Semi-Supervised AB/BA (ssAB/BA) analysis is a technique to estimate rFPR and rRecall metrics, while avoiding the need to label utterances by human annotation, which are instead estimated in a semi-supervised way. Because of that, the technique has lower cost and is faster to run than AB/BA analysis. The process also improves audio privacy, since no audio is listened to by annotators. However, since it relies on a machine-labeling process, the technique is more susceptible to errors due to bias.

There are several lines of research on Semi-Supervised Learning models, such as Teacher models (Li et al., 2017; Tarvainen and Valpola, 2017), which provide posterior probabilities as soft labels in order to train Student models.

Assuming that we have a Label Machine $\mathcal{M}$, we apply this machine on utterance i to produce a score m_i. However, if m_i produces a soft label m_i, a mapping function $\phi_{\mathcal{M}}(m_i) = p_i$ can be used to convert the machine-generated score m_i from machine $\mathcal{M}$ to a probability of true accept p_i. Appendix A.4 illustrates this process using a polynomial mapping.

When only soft labels are available, representing a probability of true label, we can apply the Bayes' theorem on Equation (1) to calculate rFPR using:

$$rFPR = \frac{\frac{P(L=0|B=1,A=1)*P(B=1|A=1)}{P(L=0|A=1)}}{\frac{P(L=0|A=1,B=1)*P(A=1|B=1)}{P(L=0|B=1)}}, \quad (7)$$

3

which results in:

$$rFPR = \frac{\sum_{i=0}^{N_A} p(L_i = 0|A_i = 1, B_i = 1)}{\sum_{i=0}^{N_A} p(L_i = 0|A_i = 1)} *$$

$$\frac{\sum_{i=0}^{N_B} p(L_i = 0|B_i = 1)}{\sum_{i=0}^{N_B} p(L_i = 0|B_i = 1, A_i = 1)},$$

(8)

Equations (3) and (8) are, therefore, equivalent if $p(L_i = 0)$ is a hard ground-truth label (either 0 or 1).

Similarly, in Semi-Supervised AB/BA the rRecall can be written as:

$$rRecall = \frac{\sum_{i=0}^{N_A} p(L_i = 1|A_i = 1, B_i = 1)}{\sum_{i=0}^{N_A} p(L_i = 1|A_i = 1)} *$$

$$\frac{\sum_{i=0}^{N_B} p(L_i = 1|B_i = 1)}{\sum_{i=0}^{N_B} p(L_i = 1|B_i = 1, A_i = 1)},$$

(9)

where $p(L_i = 1) = 1 - p(L_i = 0)$ is the probability of TP for a given utterance i.

2.3 Threshold selection

Another important aspect to consider during a KWS model evaluation is the model sensitivity. In order to determine if a given utterance will be considered an accept or reject by the model, assuming high model scores s_i are given to higher chance of detection, the utterance i will be considered an accept by the model X, i.e., $X_i = 1$, if $s_i > t_X$, where t_X is a threshold attributed to model X sensitivity.

Notice that t_X directly impacts the rFPR and rRecall. Essentially, as we increase t_X, it will make the model more restrictive, so both FPR and Recall are reduced. Appendix A.3 gives an example where the threshold t_B of the candidate model B is found according to the trade-off between rFPR and rRecall.

3 Experiments

In this section we present experiments to show the performance of AB/BA and ssAB/BA on simulation and real data.

3.1 Simulations

This section presents the simulations performed.

3.1.1 AB/BA analysis simulation

We created a simulation to show how the calculations from the AB/BA formulas (Equations (2) and (3)) are equivalent to the direct computation of rFPR and rRecall, but without the need for data not accepted by the models. We also show confidence intervals on those metrics as a function of the number of labels. Assume we have a source that emits positive utterances with a probability of 0.3. The model A has a Recall of 0.8 and FPR of 0.1. We consider two pairs of values for model B: Recall and FPR of (0.82, 0.075) leading to rRecall and rFPR of (1.025, 0.75), and Recall and FPR of (0.84, 0.05) leading to rRecall and rFPR of (1.05, 0.5). In addition, data accepted by model A has a probability of being accepted by B of 0.95 and 0.5 for TPs and FPs respectively. We run the simulation considering half of the data is collected by A, half by B. The simulation assumes that the model that does not collect the data is run only on the accepts of the model that does. We report estimated rFPR, rRecall in 3 scenarios, using AB/BA direct estimation, AB/BA with the introduced assumption Equation (6), and using a classic A/B test (only estimating rFPR in this case).

Table 1 shows the rRecall and rFPR, along with 95% confidence interval from 1000 bootstrapping replicates. We can notice in the table that we can detect improvement as small as a 5% Recall improvement and 50% FPR reduction by labeling less than 5000 utterances. We can also see that using the assumption of same TPs and FPs between the models, represented as Approx. AB/BA in the table, to estimate 25% rFPR improvement, the confidence intervals shrink from $-25\%(-52\%, +16\%)$ to $-26\%(-48\%, +2\%)$, while keeping median estimation equally accurate, showing that this is a helpful assumption. The table also shows that the rFPR predicted by regular A/B-Test on the model accepted data also gives close estimation according to the simulated parameters. However, this approach leads to higher confidence interval than the proposed approach and cannot estimate the rRecall.

3.1.2 Semi-Supervised AB/BA analysis simulation

One important point when working with ssAB/BA is with respect to the quality of the label generation process. Therefore, we start with a simulation showing this effect.

In our experiment, we do a Monte Carlo simulation by generating data that can be used to compute

4

Streams	Labeled	Expected rRecall / rFPR Improv.	rRecall / rFPR Direct AB/BA	rRecall / rFPR Approx. AB/BA	rFPR AB-Test
10K	500	1.025 / 0.75	1.025 [0.963, 1.103] / 0.75 [0.48, 1.16]	1.025 [0.964, 1.102] / 0.74 [0.52, 1.02]	0.74 [0.43, 1.19]
100K	5K	1.05 / 0.5	1.051 [1.028, 1.075] / 0.5 [0.45, 0.55]	1.051 [1.028, 1.075] / 0.49 [0.45, 0.54]	0.49 [0.41, 0.58]

Table 1: **Simulation of AB/BA analysis**

the rRecall and rFPR metrics. Data is generated such that models A and B collect, respectively, 40% and 20% as TPs. The probability that FPs and TPs from model A are also accepted by model B are, respectively, 0.3 and 0.9, and the probability that FPs and TPs from model B are also accepted by model A are, respectively, 0.6, and 0.8. Then, we simulate three soft label machines $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ using a Beta distribution.

Among the three machines, $\mathcal{M}_1$ is simulated to generate soft-labels with $\mathcal{B}(2, 1000|L = 0)$ and $\mathcal{B}(300, 5|L = 1)$ to have the same accuracy for both models A and B that will be evaluated in ssAB/BA. Then, $\mathcal{M}_2$ is simulated with $\mathcal{B}(5, 100|B = 1, L = 0)$ and $\mathcal{B}(300, 5|B = 1, L = 1)$ to make more mistakes in the form of higher TP probability on the FPs collected by model B, and $\mathcal{M}_3$ with $\mathcal{B}(2, 1000|B = 1, L = 0)$ and $\mathcal{B}(100, 10|B = 1, L = 1)$ to make more mistakes in the form of lower TP probability on the TPs collected by the model B. Notice that we keep the accuracy of the machines on model A data constant, since B is a candidate model with new data never seen before.

Notice that based on the parameters chosen the expected AB/BA rRecall for the simulation is 1.12, since $P(B = 1|A = 1, L = 1)/P(A = 1|B = 1, L = 1) = 0.9/0.8 = 1.12$. Similarly, the expected rFPR is 0.5. Results of the simulation, along with 95% confidence interval from 1000 bootstrapping replicates, are shown in Table 2.

	rRecall	rFPR
AB/BA	1.12[1.10,1.14]	0.50[0.48,0.52]
ssAB/BA $\mathcal{M}_1$	1.12[1.10,1.13]	0.51[0.49,0.53]
ssAB/BA $\mathcal{M}_2$	1.08[1.05,1.10]	0.51[0.49,0.53]
ssAB/BA $\mathcal{M}_3$	1.12[1.10,1.15]	0.56[0.54,0.57]

Table 2: **Simulation comparing AB/BA and ssAB/BA analysis according to label quality**

From Table 2, as expected, $\mathcal{M}_1$ results are almost exactly the same as in AB/BA, since $\mathcal{M}_1$ mean TP probabilities are 0.2% for $L = 0$ and 98.4% for $L = 1$, which are close to ground-truth. When using $\mathcal{M}_2$, however, we can see that the

rRecall measured drops from 1.12 to 1.08, as this model makes more mistakes on the FPs collected from model B, increasing the machine TP probability on this data from 0.2% to 4.8%. In this case, we can see that the rFPR is unchanged at the reported precision. Similarly, in the case of $\mathcal{M}_3$ we see a change in the reported rFPR, which increases from 0.51 to 0.56 as this model makes more mistakes on the TPs collected from model B, dropping the TP probability on this data from 98.4% to 90.9%. In this case, the rRecall is mostly unchanged. This results show that, although the label-generation process by machine can be imperfect, it gives good approximations on the Recall Ratio and FPR Ratio in order to make deployment decisions.

3.2 AB/BA Analysis on Real Application

Next, we show how AB/BA performs when applied to real customer data in order to guide the decision on how much customer experience is being improved with the deployment of new KWS models. Comparison between AB/BA and ssAB/BA analysis on real data is show in Appendix A.5.

3.2.1 Comparison between deployments with different threshold t_b

We show results from two real deployments, called D_1 and D_2. In D_1, the AB/BA analysis ratio metrics are used to compare a baseline model to a candidate model with high threshold (more restrictive), while in D_2 the same baseline is compared to the same candidate model with low threshold (more permissive). The two deployments use about 5000 annotated utterances per model. Results are show in Table 3.

As we can see in Table 3, D_1 resulted in loss of Recall by 5% relative, but improving the rFPR in 43% relative. By deploying the candidate model

D	rRecall	rFPR
D_1	0.95 [0.94-0.96]	0.57 [0.46-0.68]
D_2	1.07 [1.05-1.09]	1.33 [1.17 - 1.44]

Table 3: **AB/BA Analysis results when deploying models with different thresholds**

Dataset	rRecall	rFPR
Test set	1.03	0.5
AB/BA	1.16[1.15-1.19]	0.21[0.14-0.26]

Table 4: **AB/BA Analysis Recall Ratio comparison to test set metrics**

with low t_B threshold (D_2), the AB/BA analysis then shows $7\%(5\%, 9\%)$ relative improvement in Recall, with a trade-off of $33\%(17\%, 44\%)$ relative increase in FPR. AB/BA analysis correctly found that the D_1 model is a more conservative model, and D_2 a more sensitive model than the baseline model.

3.2.2 Comparison between AB/BA analysis and the evaluation a test dataset

New candidate models are evaluated on offline test datasets to decide if they can be deployed to customers. However, offline test datasets are composed of utterances collected by the current or older deployed models, and recall improvement of those candidate models may be under-estimated. Here we show measurements from offline evaluation and AB/BA analysis in a real deployment. The AB/BA analysis was performed with approximately 7000 annotated utterances per model.

As we can see in Table 4, the rRecall estimation from the offline test set resulted in 3% relative improvement at the same FPR. That was significantly less than the 16% relative improvement measured during AB/BA, in which data from model B is used in the analysis. Similar observation can be made in terms of rFPR, where evaluation on the test set resulted in a 50% relative improvement at the same Recall, but by also accounting for data collected by the B model in AB/BA, we see that the improvements was 79% relative. This shows the importance of techniques such as AB/BA analysis to better assess the customer experience on the model being deployed.

4 Conclusion

In this paper, we presented a new framework called AB/BA analysis for recall improvement estimation of KWS under high privacy settings. We have shown that by running a candidate model offline on data collected by a baseline, and the baseline model offline on data collected by the candidate model, we were able to compute the relative Recall and relative FPR ratios using only utterances accepted by both models and use it to indicate if a candidate

model is better than the baseline. We have also shown that reasonable assumptions can be used to construct an estimator with low variance, even when the number of FPs is small. Finally, we saw that a Semi-Supervised formulation of AB/BA can be used with machine-generated labels representing the probability of a true accept. This techniques brings further improvements to AB/BA analysis, especially regarding privacy on the audio collected, which does not need to be listened to and annotated.

In the past few years, much improvement has been made on the KWS systems. Evaluation metrics, however, are typically based on previously collected data from similar KWS models or data from other sources, resulting in evaluations that do not necessarily translate to customer experience, as it fails to show improvements in data never collected due to privacy constraints. Given that not much attention has been paid to this research topic, we believe that AB/BA analysis is a valuable contribution, and we hope it helps bringing interest in this line of research.

References

Richard E Barlow and Hugh D Brunk. 1972. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147.

Guoguo Chen, Carolina Parada, and Georg Heigold. 2014. Small-footprint keyword spotting using deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, pages 4087–4091. IEEE.

Amit S. Chhetri, Philip Hilmes, Trausti Kristjansson, Wai Chu, Mohamed Mansour, Xiaoxue Li, and Xianxian Zhang. 2018. Multichannel audio front-end for far-field automatic speech recognition. In *26th European Signal Processing Conference, EUSIPCO 2018, Roma, Italy, September 3-7, 2018*, pages 1527–1531. IEEE.

Paul HC Eilers and Brian D Marx. 1996. Flexible smoothing with b-splines and penalties. *Statistical science*, 11(2):89–121.

Yixin Gao, Noah D. Stein, Chieh-Chi Kao, Yunliang Cai, Ming Sun, Tao Zhang, and Shiv Naga Prasad Vitaladevuni. 2020. On front-end gain invariant modeling for wake word spotting. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 991–995. ISCA.

Jinyu Li, Michael L Seltzer, Xi Wang, Rui Zhao, and Yifan Gong. 2017. Large-scale domain adap-

tation via teacher-student learning. *arXiv preprint arXiv:1708.05466*.

Iván López-Espejo, Zheng-Hua Tan, John Hansen, and Jesper Jensen. 2021. Deep spoken keyword spotting: An overview. *CoRR*, abs/2111.10592.

Benjamin A. Miller, Jeremy Vila, Malina Kirn, and Joseph R. Zipkin. 2018. Classifier performance estimation with unbalanced, partially labeled data. In *International Workshop on Cost-Sensitive Learning, COST@SDM 2018, San Diego, California, USA, May 5, 2018*, volume 88 of *Proceedings of Machine Learning Research*, pages 4–16. PMLR.

Anirudh Raju, Sankaran Panchapagesan, Xing Liu, Arindam Mandal, and Nikko Strom. 2018. Data augmentation for robust keyword spotting under playback interference. *CoRR*, abs/1808.00563.

Tara N. Sainath and Carolina Parada. 2015. Convolutional neural networks for small-footprint keyword spotting. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 1478–1482. ISCA.

Changhao Shan, Junbo Zhang, Yujun Wang, and Lei Xie. 2018. Attention-based end-to-end models for small-footprint keyword spotting. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 2037–2041. ISCA.

Michael Stolberg. 2006. Inventing the randomized double-blind trial: the nuremberg salt test of 1835. *Journal of the Royal Society of Medicine*, 99(12):642–643.

Ming Sun, David Snyder, Yixin Gao, Varun K. Nagaraja, Mike Rodehorst, Sankaran Panchapagesan, Nikko Strom, Spyros Matsoukas, and Shiv Vitaladevuni. 2017. Compressed time delay neural network for small-footprint keyword spotting. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 3607–3611. ISCA.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*.

George Tucker, Minhua Wu, Ming Sun, Sankaran Panchapagesan, Gengshen Fu, and Shiv Vitaladevuni. 2016. Model compression applied to small-footprint keyword spotting. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 1878–1882. ISCA.

Minhua Wu, Sankaran Panchapagesan, Ming Sun, Jiacheng Gu, Ryan Thomas, Shiv Naga Prasad Vitaladevuni, Björn Hoffmeister, and Arindam Mandal. 2018. Monophone-based background modeling for two-stage on-device wake word detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 5494–5498. IEEE.

A Appendix

A.1 Basic Evaluation Concepts

During our discussion we assume that we have a classification model, $\mathcal{M}(x) \to y$, which gets an arbitrary input x and output $y \in \{0, 1\}$, where 1 and 0 represent, respectively, a positive and negative label.

A.1.1 Precision and Recall

The Precision and Recall metrics help to distinguish between Type-I and Type-II errors. They are defined as the following:

$$Precision = \frac{TP}{TP + FP} \qquad (10)$$

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{P} \qquad (11)$$

Precision measures the proportion of correct positive predictions with respect to everything the model believes is a positive, where lower precision represents more Type-I errors. On the other hand, Recall measures the proportion of correct positive predictions with respect to all the data that is actually positive, where lower recall represents more Type-II errors.

As illustration, in a case where 90% of the data has label 1 and the model always gives output 1, this model will have 100% recall, and 90% precision.

A.1.2 False Positive Rate and False Discovery Rate

There are also multiple ways to evaluate a model with respect to the number of False Positives (FPs) it makes. Two of them, which are explored in this paper, are False Positive Rate (FPR) and False Discovery Rate (FDR). They are defined as:

$$FPR = \frac{FP}{N} \qquad (12)$$

$$FDR = 1 - Precision = \frac{FP}{TP + FP} \qquad (13)$$

Therefore, we can see that the FPR is similar to Recall in the sense that the reference is only one class of the data, which in this case are the negative samples. It measures the proportion of negative samples that are miss classified by a model. The FDR, however, is the complement to the Precision, measuring the number of false positives among all samples that the model classifies as positive.

Considering again the example where 90% of the data has label 1 and the model always gives output 1, its FPR is 100%, while its FDR is 10%.

A.1.3 A/B Test

Another related concept to our proposed method is the A/B test. The A/B test is a frequently used technique in multiple areas, such as medicine (Stolberg, 2006), marketing, political campaigning, product pricing, among others. The technique consists of doing a hypothesis test by giving two randomized and unbiased set of populations, called A and B, two different versions of the subject being compared. For example, in the context of marketing, one could choose to give two different versions of user interface to users in order to measure differences in engagement, which is measured through statistical tests.

In the context of model evaluation, the A/B test can be explored by given two different models to users. After data collection, metrics can be computed for each population, such as metrics presented in Section A.1, and statistical tests, such as t-test, can be used to compare the metrics in the different populations.

The A/B test has, however limitations, when used to evaluate keyword spotting with audio privacy settings. Given that only data where the keyword is detected is collected by the models, the amount of negative data is highly biased towards the false positives from the models that collected the data. Therefore, metrics that rely on negative data, such as Recall and FPR, cannot be computed and compared using A/B test. This limitation is explored in this paper with our proposed AB/BA technique in order to tackle this challenge of Recall improvement estimation.

A.2 Stratified Sampling

To decide if a model is better than the other one, we could simply annotate the utterances where the two models disagree. However, if it is also desirable to calculate absolute metrics, such as False Positive Rate (FPR), then annotation of model agreements is also needed. In order to reduce the amount of annotations for this task, we propose to use stratified sampling, with two strata, agreement and disagreement, such that different number of annotations are done per strata.

Our stratified sampling strategy uses the Neyman allocation principle. The optimal number of

annotations N_j for a strata j can be found using:

$$N_j^* = \frac{Nw_j\sqrt{p_j(1-p_j)}}{\sum_{i=1}^{L} w_i\sqrt{p_i(1-p_i)}}, \qquad (14)$$

where N is our annotation budget, w_j is the proportion in each strata, and p_j is the expected model FPR, assumed to be estimated, for example, from previously annotated data.

The efficiency improvement of the stratified sampling strategy compared to a random sampling strategy, in terms of variance in the FPR, is:

$$
\begin{aligned}
Eff &= 1 - \frac{V(fpr^*)}{V(fpr)} \\
&= 1 - \frac{\frac{1}{N}(\sum_j w_j \sqrt{p_j(1-p_j)})^2}{\frac{1}{N}p(1-p)},
\end{aligned}
\qquad (15)
$$

For the purpose of illustration, assume we have a 10% disagreement between models, with FPR of 20% in this strata, whereas the agreement strata has a FPR of 5%, and the overall FPR of 8%, Equation (14) gives us, for the disagreement strata:

$$
\begin{aligned}
\frac{N_j^*}{N} &= \frac{0.1\sqrt{0.2(1-0.2)}}{0.1\sqrt{0.2(1-0.2)} + 0.9\sqrt{0.05(1-0.05)}} \\
&\approx 17\%,
\end{aligned}
\qquad (16)
$$

indicating that the disagreement strata should optimally be 17% of the annotation budget, without affecting the FPR variance. The efficiency gain of this method is:

$$
\begin{aligned}
Eff &= \\
&1 - \frac{(0.1\sqrt{0.2(1-0.2)} + 0.9\sqrt{0.05(1-0.05)})^2}{0.08(1-0.08)} \\
&\approx 24\%,
\end{aligned}
\qquad (17)
$$

indicating that the annotation budget can be reduced by approximately 24%, without affecting the overall FPR variance.

A.3 Threshold Selection Example

Given two models A and B, where A is a baseline and B a candidate, we assume that A has a known t_A, previously obtained according to the desired FPR and Recall trade-off. Therefore, in order to determine if the candidate model B is better, we

can calculate the FPR Ratio and Recall Ratio for multiple thresholds t_B. Next, the decision will be guided towards the goal of model B. Say, for example, that the goal of model B is not to improve FPR, but only Recall, then we select $t_B = t$ such that when applying this threshold it results in $FPR\_Ratio = 1$. One example, illustrating this process, is given in Table 5.

t_B	FPR Ratio	Recall Ratio
0.1	1.5	1.20
0.2	1.0	1.05
0.3	0.8	1.01
0.4	0.7	0.98

Table 5: **Threshold selection:** The table shows an example of how FPR Ratio and Recall Ratio change, as a function of t_B. It shows that, by selecting $t_B = 0.2$, model B has the same FPR Ratio as model B, but improves Recall in 5%.
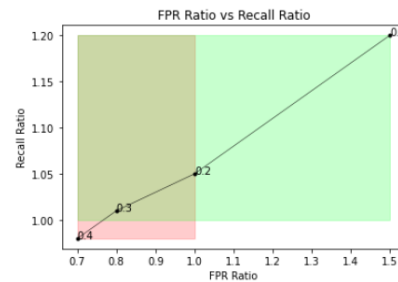

Figure 2: **Model sensitivity selection:** The green area in the figure shows the region where the recall is improved, and the red area where the FPR improves. We can see that 0.2 and 0.3 are potential thresholds for t_B to improve the performance of model A.

Table 5 shows the FPR Ratio and Recall Ratio as a function of t_B, similar to the data behind a traditional DET curve, but using the proposed ratio metrics. We can see that, by selecting $t_B = 0.2$, model B has the same FPR Ratio as model B, but improves Recall in 5% relative. It is also interesting to see that using $t_B = 0.3$ causes improvement in both FPR Ratio and Recall Ratio. Using $t_B = 0.1$ has 20% relative improvement in Recall, but with a high trade-off in FPR and, similarly $t_B = 0.4$ has 30% relative improvement in FPR, but with degradation in Recall.

We see, therefore, that the model threshold can be selected according to the goal in the trade-off between FPR Ratio and Recall Ratio. It is clear though that the model B is superior than model A, since it has a threshold region that improves both FPR and Recall, as shown in Figure 2.

A.4 Soft-Label Score Calibration

Given a set of utterances composed of N machine-generated soft labels $m \in R^N$ and human labels $y \in R^N$, we propose to learn $\phi(m_i)$ using a polynomial of degree 3. Notice that, as the target variable is binary, and we expect to have more TAs as m increases, the polynomial should be a monotonic increasing function between $[0, 1]$. Although not guaranteed, our experiments show this to be an empirically good choice. One example is shown in Figure 3. However, since the polynomial is not guaranteed to have probabilities bounded between $[0, 1]$, we have also explored to use a B-Spline Logistic Regression with monotonic constraints (Eilers and Marx, 1996; Barlow and Brunk, 1972). However we have not seen significant difference and have decided to use the polynomial approach for simplicity.

Once we have soft-labels for the utterances from model A and B to be compared, we can use Equations (8) and (9) in order to estimate the FPR Ratio and Recall Ratio trade-off.
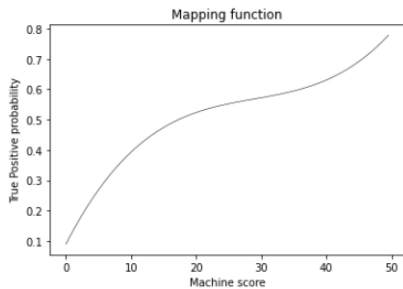

Figure 3: **Probability mapping function:** Using a polynomial of degree 3, arbitrary for this illustration, we can see that scores from a model between 0 and 50 are mapped to a probability of True Positive.

A.5 Semi-Supervised AB/BA Comparison to AB/BA on Real Data

Next we show real data examples where we used both AB/BA and ssAB/BA, in order to see if the ratios reported are similar. In this case, machine-generated scores were generated by a cloud-side verification system, and its scores converted to a TP probability according to a polynomial mapping function learned from other existing labeled datasets, following the process described in Section 2.2. Results are shown in Table 6.

Results in Table 6 show that ssAB/BA is able to well-approximate the AB/BA results, having results with overlapping margin of error in most

Example	ssAB/BA rRecall at rFPR	AB/BA rFPR at rRecall
1	1.03 [1.03-1.03]	1.04[1.04-1.05]
2	1.01 [1.01-1.01]	1.0[0.97-1.05]
3	0.99 [0.98-1.01]	1.01[0.99-1.04]

Table 6: **Semi-Supervised AB/BA and AB/BA analysis comparison on real data:** The table shows three examples comparing AB/BA and Semi-Supervised AB/BA on real customer data

cases. It is important to notice, however, that there is a risk of using ssAB/BA related to the quality of labels generated. In the Example 3, we can see that ssAB/BA results suggest a Recall loss of 1% relative, while AB/BA suggests a Recall improvement of 1% relative, although the confidence intervals overlap. That represents the case where, when the recall improvement is small, the uncertainty of the machine-label generation may limit its applicability. It is important, therefore, to monitor the quality of the label machines in order to know how trustworthy they are, and to also use other auxiliary metrics that help reducing the risk of trusting the ssAB/BA results by itself.