Speaker-adapted confidence measures for speech recognition of video lectures

Isaias Sanchez-Cortina, Jesús Andrés-Ferrer, Alberto Sanchis, Alfons Juan

MLLP, DSIC, Universitat Politècnica de València.
Camí de Vera, s/n, 46022 València, Spain.

Abstract

Automatic Speech Recognition applications can benefit from a *confidence measure* (CM) to predict the reliability of the output. Previous works showed that a *word-dependent naïve Bayes* (NB) classifier outperforms the conventional word posterior probability as a CM. However, a discriminative formulation usually renders improved performance due to the available training techniques.

Taking this into account, we propose a *logistic regression* (LR) classifier defined with simple input functions to approximate to the NB behaviour. Additionally, as a main contribution, we propose to adapt the CM to the speaker in cases in which it is possible to identify the speakers, such as online lecture repositories.

The experiments have shown that speaker-adapted models outperform their non-adapted counterparts on two difficult tasks from English (videoLectures.net) and Spanish (poliMedia) educational lectures. They have also shown that the NB model is clearly superseded by the proposed LR classifier.

Keywords: confidence measures, speech recognition, speaker adaptation, log-linear models, online video lectures

1. Introduction

Significant advances in the field of *Automatic Speech Recognition* (ASR) have been achieved over the last decades. Nowadays, automatic transcriptions of spontaneous speech in moderately noisy environments have reached an accurate enough quality ([1, 2, 3]). This quality can be even better when ASR systems are adapted to specific scenarios ([4, 5, 6, 7, 8, 9]). Nonetheless, ASR is still far from producing error-free transcriptions and, consequently, its performance in many applications is not completely satisfactory.

To further improve the usefulness and performance of the current technology, researchers have proposed to compute a normalised score or *confidence measure* (CM) to indicate the reliability of the ASR output. This score has been computed at different levels: phoneme, word, phrase or sentence. Nevertheless, CM at the word level has been the main focus in the literature due to its usefulness for the vast majority of applications ([10, 11, 12, 13, 14, 15, 16]).

Email address: `{isanchez,jandres,josanna,ajuan}@dsic.upv.es` (Isaias Sanchez-Cortina, Jesús Andrés-Ferrer, Alberto Sanchis, Alfons Juan)

One widely used word-level CM has been word posterior probability ([10]). From then on, many works have focused on combining word posterior with additional sources of knowledge. The combination has been addressed as a classification problem in the vast majority of the works. Most well-known classifier algorithms have been tried: linear, Gaussian mixtures, neural networks, decision trees, support vector machines, etc. For further reference, a still good comprehensive survey can be found in [17].

In the framework of CM as a classification problem, significant improvements were achieved by means of a combination of word-dependent (specific) and word-independent (generalised) *naïve Bayes* (NB) classifiers [18]. Nonetheless, NB is learned by means of a generative criterion, the *maximum likelihood estimate* (MLE), which involves some issues. In particular, MLE overfits due to the unseen data. This issue was addressed in NB work by using a complex *backing-off* smoothing technique. But still, MLE aims at modelling the distribution underlying a given sample, which does not guarantee the solution to be the best suited for classification. Indeed, better fitted criteria may improve overall performance. For instance, the *maximum mutual information* (MMI) [19] aims at better discriminating between classes without explaining the data. This criterion has been widely exploited in the literature for the *maximum entropy* (ME) models ([20, 21]).

Nevertheless, despite the success of MMI training in many applications, there is no direct relationship between maximising the MMI and minimising the probability of classification error. Instead, there are better suited criteria, which guarantee the minimisation of the *classification error rate* (CER) such as the *minimum classification error* (MCE) or the *mean squared error* (MSE). Therefore, we propose a *logistic regression* (LR) model to be learnt by means of the MSE to surpass NB performance.

On the other hand, speaker model adaptation has proved to be very effective for the improvement of recognition performance [4, 5, 6, 7]. However, adaptation of the CMs to the speaker is nowadays unexplored. There is an increasing number of interesting scenarios in which CMs can be very useful and information about the speaker is available, such as the online lecture repositories. These repositories usually count with a large number of speeches delivered by a reduced number of speakers. Improving CM performance in these academic repositories is highly motivated since manual transcription is not affordable for such a large amount of speeches. Moreover, ASR performance is usually poor due to the amount of technical concepts, very different native and non-native accents, etc. In this scenario, *interactive speech transcription* (IST) guided by CMs can help in massively producing acceptable transcripts for large amounts of videos with limited manual effort [22].

Motivated by the scenario depicted above, we propose to adapt the CM models to the speaker in an attempt to improve CM classification and IST performance. To do so, we formulate the speaker adaptation to extend both the published NB and the proposed LR models.

The rest of the content is organised as follows: the inclusion of speaker dependence into the NB model is described in Section 2. Section 3 proposes the LR model and formulates its corresponding speaker-dependent version. Section 4 describes the evaluation of the proposed models on two challenging tasks based on ASR transcripts from videoLectures.net and poliMedia repositories. Comparative results are presented including also Conditional Random Field (CRF) models [23, 24, 25]. Section 5 proves that the increased CM performance results in better amended transcripts for videoLectures.net when integrated into an IST application. Finally, Section 6 raises the conclusions.

2. Speaker-adapted naïve Bayes classifier

In this Section, we introduce a speaker-adapted confidence estimator model. The model is designed to extend the *naïve Bayes* (NB) approach that was successfully applied to speech recognition [18] as well as to machine translation [26]. Thus, let us first briefly recall the speaker independent NB model.

The NB model is formulated on the framework of confidence estimation addressed as a classification problem. On this framework, the recognised words are labelled as correct ($c = 1$) or incorrect ($c = 0$) by means of the class posterior given the word (w) and a vector of input scores ($\mathbf{x}$):

$$\hat{c} = \arg\max_{c} \, p(c|w, \mathbf{x}) = \arg\max_{c} \, p(c|w) \, p(\mathbf{x}|c, w) \tag{1}$$

where Eq. 1 is obtained by applying Bayes' rule and then ignoring the class-independent term. Also, the values of all the involved variables in the latter equation are assumed to be discrete. Discretisation avoids the need of explicitly modelling the probability distribution of continuous-valued features, while it renders a more flexible and data-driven model. Details on discretisation and several different approaches can be found in [23]. Here, we just discretised by dividing the feature domain into a fixed number of evenly-spaced bins. The optimal number of bins was tested on the development set.

The estimation of $p(\mathbf{x}|c, w)$ is usually biased due to the training data sparsity. More robust estimations can be obtained by simplifying the problem with the following strong independence assumption (the "naïve Bayes assumption"):

$$p(\mathbf{x}|c, w) = \prod_{d}^{D} p(x_d|c, w) \tag{2}$$

Therefore, the basic problem is to estimate $p(x_d|c, w)$ for each class-word pair and $p(c|w)$ for each target word. Given N training samples $\{(\mathbf{x}_n, c_n, w_n)\}_{n=1}^{N}$, these probabilities can be computed as the *maximum likelihood estimate* (MLE):

$$p(c|w) = \frac{N(c, w)}{N(w)} \qquad p(x_d|c, w) = \frac{N(c, w, x_d)}{N(c, w)} \tag{3}$$

where $\{N(\cdot)\}$ are suitably defined event counts on a given training data set. However, the MLE quickly overfit the training data. In order to prevent this overfitting, a particular backing-off smoothing method was introduced in [18].

We propose to extend the NB into a *naïve Bayes speaker-adapted* model (NB+spk). For that, a new variable s is introduced into Eq. 1 to identify the speaker:

$$\hat{c} = \arg\max_{c} \, p(c|w, \mathbf{x}, s) = \arg\max_{c} \, p(c|w, s)p(\mathbf{x}|c, w, s) \tag{4}$$

Consequently, the problem is turned into computing $p(x_d|c, w, s)$ for each class-word-speaker triplet and $p(c|w, s)$ for each word-speaker pair, which analogously can be simply estimated by means of their MLE:

$$\hat{p}(c|w,s) = \frac{N(c,w,s)}{N(w,s)} \tag{5}$$

$$\hat{p}(x_d|c,w,s) = \frac{N(c,w,x_d,s)}{N(c,w,s)} \tag{6}$$

As in the non-adapted (speaker-independent) NB approach, MLE overfitting can be prevented by using a straightforward extension of the backing-off smoothing method proposed in [18].

3. Speaker-adapted logistic regression confidence estimator

In this Section, we propose a new CM model based on *logistic regression* (LR) models. It is worth noting that, for binary classification problems such as the CM problem considered in this work, these models are equivalent to more general *conditional random field (CRF)* models. In what follows, after describing the proposed speaker-dependent LR models, we briefly discuss how to discriminatively learn them using the MSE training criterion (Sec. 3.1). In Sec. 4, this criterion is empirically compared with a similar yet different criterion that is commonly used in CRF training.

The proposed approach resembles the ones presented in [27]. However, that work formulated the classification problem as a generative model, and only the posterior of the features was attempted to be learnt in a discriminative way. Furthermore, the purpose was to mimic NB, so no improvements were obtained. Hence, in contrast to [27], here we do model the class posterior; define simpler input functions for the LR model; introduce a standard L_2 regularisation to avoid the complex set of maximum entropy constraints with cut-offs; and use the MSE learning criterion, optimised with the simple and fast *iRPROP+* [28] algorithm.

The assumption of a general LR distribution for the class posterior yields the following classification rule:

$$\hat{c} = \arg\max_{c} p(c|w,\mathbf{x}) = \arg\max_{c} \frac{\exp\left(\sum_i \lambda_i f_i(c,w,\mathbf{x})\right)}{\mathcal{Z}(w,\mathbf{x})} \tag{7}$$

where w is the recognised word and $\mathbf{x} = (x_1..x_D)$ is a D-dimensional vector of discretised input features. On the other hand, $\mathcal{Z}$ is a normalisation constant, which does not affect classification; $\lambda_{(\cdot)}$ are a set of data-driven parameters; and $f_{(\cdot)}$ a set of functions which give the model expressiveness.

As discussed before, the NB model in [18] introduced several convenient assumptions: conditional independence amongst the D scores, discretisation of the continuous-valued scores, etc. Hence, we propose now a particular definition for the $f_{(\cdot)}$ functions to make the LR model behave similarly to the NB model in terms of classification.

Let i be the triplet of labels ($\tilde{c} \in \{0,1\}$, $\tilde{w} \in \{1..W\}$, $\tilde{x}_{\tilde{d}} \in \{1..X_{\tilde{d}}\}$) indexing the classes, the known vocabulary and the values of the score number $\tilde{d} \in \{1..D\}$

respectively. $X_{\tilde{d}}$ accounts for the total number of different possible discrete values of $x_{\tilde{d}}$. For each possible triplet, let us define the following function:

$$f_{\tilde{c},\tilde{w},\tilde{x}_{\tilde{d}}}(c,w,\mathbf{x}) = \quad \delta_{\tilde{c}}(c)\cdot\delta_{\tilde{w}}(w)\cdot\delta_{\tilde{x}_{\tilde{d}}}(\mathbf{x}) \tag{8}$$

with $\delta(\cdot)$ being the Kronecker delta and $\delta_{\tilde{x}_{\tilde{d}}}(\mathbf{x}) \equiv \prod_{d'}^{D} \delta_{\tilde{x}_{\tilde{d}}}(x_{d'})\cdot\delta_{\tilde{d}}(d') = \delta_{\tilde{x}_{\tilde{d}}}(x_{\tilde{d}})$.

It becomes clear from the latter definition that the set of functions $\{f_{\tilde{c},\tilde{w},\tilde{x}_{\tilde{d}}}\}$ serves merely to activate the corresponding weights $\{\lambda_{\tilde{c},\tilde{w},\tilde{x}_{\tilde{d}}}\}$. Thus, it is the set of weights alone which will render the classification, and they are to be learned exclusively from data, as detailed in Sec 3.1. Also, it should be noted that each of the defined functions does not involve more than one score. This is precisely equivalent to assuming naïve Bayes over the scores, as in Eq. 2.

Furthermore, in order to prevent overfitting, additional weights and functions to be active independently of one or more label values are necessary:

$$\begin{aligned}
f_{\tilde{c},\emptyset,\emptyset}(c,w,\mathbf{x}) &= \quad \delta_{\tilde{c}}(c) \\
f_{\tilde{c},\emptyset,\tilde{x}_{\tilde{d}}}(c,w,\mathbf{x}) &= \quad \delta_{\tilde{c}}(c)\cdot\delta_{\tilde{x}_{\tilde{d}}}(\mathbf{x}) \\
f_{\tilde{c},\tilde{w},\emptyset}(c,w,\mathbf{x}) &= \quad \delta_{\tilde{c}}(c)\cdot\delta_{\tilde{w}}(w)
\end{aligned} \tag{9}$$

These terms enable a behaviour similar to the smoothing in the NB model, which backs off to less specific probabilities under certain conditions.

Finally, it should be noted that the presented model typically involves a huge number of weights to be estimated, of order $\mathcal{O}$(vocabulary × number of features × mean number of values per score). Fortunately, the computation time can be halved by defining a new set of weights $\lambda_{(\cdot)} \equiv \lambda_{\tilde{c}=1,(\cdot)} - \lambda_{\tilde{c}=0,(\cdot)}$, and the corresponding activation features:

$$\begin{aligned}
f_{\tilde{w},\tilde{x}_{\tilde{d}}}(w,\mathbf{x}) &= \delta_{\tilde{w}}(w)\cdot\delta_{\tilde{x}_{\tilde{d}}}(\mathbf{x}) \\
f_{\emptyset,\emptyset}(w,\mathbf{x}) &= 1 \\
f_{\emptyset,\tilde{x}_{\tilde{d}}}(w,\mathbf{x}) &= \delta_{\tilde{x}_{\tilde{d}}}(\mathbf{x}) \\
f_{\tilde{w},\emptyset}(w,\mathbf{x}) &= \delta_{\tilde{w}}(w)
\end{aligned} \tag{10}$$

in this way, Eq. (7) adopts the following expression:

$$p(c|w,\mathbf{x}) = \frac{1}{1+\exp\left((-1)^{c}\sum_i \lambda_i f_i(w,\mathbf{x})\right)} \tag{11}$$

Speaker dependence can be easily introduced into Eq. (11), yielding a *logistic regression speaker-adapted* (LR+spk) model:

$$p(c \mid w,\mathbf{x},s) = \frac{1}{1+\exp\Big((-1)^{c}\cdot\Big(\sum_i \lambda_i f_i(w,\mathbf{x}) + \sum_j \lambda_j f_j(w,\mathbf{x},s)\Big)\Big)} \tag{12}$$

where speaker dependence has been formulated as a separated sum over j for the sake of clarity. Now, the number of weights to be estimated is increased by S times, S being the number of known speakers. In this case, the new index j should map the triplet of labels ($\tilde{w} \in \{\emptyset, 1..W\}$, $\tilde{x}_{\tilde{d}} \in \{\emptyset, 1..X_{\tilde{d}}\}$, $\tilde{s} \in \{1..S\}$).

Thus, speaker adaptation results in the addition of the following functions:

$$\begin{aligned}
f_{\tilde{w},\tilde{x}_{\tilde{d}},\tilde{s}}(w,\mathbf{x},s) = &\quad \delta_{\tilde{w}}(w)\cdot\delta_{\tilde{x}_{\tilde{d}}}(\mathbf{x})\cdot\delta_{\tilde{s}}(s)\\
f_{\tilde{w},\emptyset,\tilde{s}}(w,\mathbf{x},s) = &\quad \delta_{\tilde{w}}(w)\cdot\delta_{\tilde{s}}(s)\\
f_{\emptyset,\tilde{x}_{\tilde{d}},\tilde{s}}(w,\mathbf{x},s) = &\quad \delta_{\tilde{x}_{\tilde{d}}}(\mathbf{x})\cdot\delta_{\tilde{s}}(s)\\
f_{\emptyset,\emptyset,\tilde{s}}(w,\mathbf{x},s) = &\quad \delta_{\tilde{s}}(s)
\end{aligned} \tag{13}$$

3.1. Discriminative learning

As discussed in Sec. 1, the weights of the discriminative models can be estimated to minimise the MSE, which may be preferable for classification problems instead of the MMI criterion and the MLE criterion for generative models. Given N training samples $\{(\mathbf{x}_n, c_n, w_n)\}_{n=1}^{N}$, the MSE can be formulated as an optimisation problem by means of the objective:

$$F_{\text{MSE}}(\lambda) = \sum_{n=1}^{N} \left(c_n - p_\lambda(c_n = 1 \mid w_n, \mathbf{x}_n)\right)^2 \tag{14}$$

However, there is no closed form solution for the optimal λ under the minimum MSE constrain. Fortunately, any simple gradient descent based optimisation algorithm can succeed in finding the solution despite the MSE not being a convex criterion. In this work we opted for the simpler iRPROP+ [28] iterative algorithm, which provides faster convergence than other more expensive methods such as *generalised iterative scaling* (GIS) [29]. A recent evaluation of different optimisation algorithms on a large task can be found in [30].

Another common issue of many training criteria, including MSE, is that they easily overfit the weights to the training data. Since there is no clear way to smooth discriminatively trained models, a typical amendment is to add a L_2 regularisation term to the objective:

$$F(\lambda) = F_{\text{MSE}}(\lambda) - \frac{C}{2}\sum_i (\lambda_i - \lambda_i^{(0)})^2 \tag{15}$$

where $\lambda^{(0)}$ can be either a reliable estimation of the weights or simply $\mathbf{0}$.

For our model, $\lambda_i^{(0)} = \mathbf{0}$ is a clever guess, since it prevents the features from having an overrated impact. During experimentation, the zero regularisation made the feature-independent term $\lambda_{\emptyset,\emptyset}$ drop quickly to zero after a few iterations. This behaviour can be interpreted as an increased generalisation of the model, since $\lambda_{\emptyset,\emptyset}$ is proportional to the logarithm of the class prior $p(c)$ from the generative point of view. Thus, for two different models yielding the same performance on a certain test, the one with $\lambda_{\emptyset,\emptyset}$ closer to zero is likely to perform better on a new test with different prior distribution.

4. Experiments

4.1. Experimental setup

The evaluation of the proposed models (NB+spk, LR and LR+spk) and the baseline model (NB) has been carried out over two difficult tasks from English (videoLectures.net) and Spanish (poliMedia) video lectures. These tasks have been used in the context of the EU-funded project transLectures, which had the aim of developing innovative, cost-effective tools for the automatic transcription and translation of online educational videos [31]. The English task has been defined over the free and open access educational video lecture repository VideoLectures.NET (VL). In VL, the recorded lectures are mostly delivered by distinguished scholars and scientists at important conferences, summer schools, workshops, etc. Currently, VL hosts more than 16.000 lectures from 12.698 speakers. The Spanish task has been defined over Polimedia (PM), which is a recent, innovative service for the creation and distribution of multimedia educational content at *Universitat Politècnica de València* (UPV). PM is designed primarily to allow UPV professors to record their courses in video blocks lasting up to 10 minutes, accompanied by time-aligned slides. PM hosts more than 9.000 lectures from 1.300 speakers with a duration of 2.100 hours.

The state-of-the-art ASR *TLK* toolkit ([32]) has been used for the experiments. Acoustic models (AM) were learned using TLK by means of a pre-trained *deep neural network hidden Markov model* (DNN-HMM) hybrid architecture, in a similar fashion to [33]. Speaker adaptation was implemented using constrained MLLR (CMLLR) features [34, 35]. The speech data to train the English AM consisted of out-of-domain corpora (TED-LIUM [36], EPPS [1, 37, 38] and Voxforge [39]), as well as in-domain VL speeches. In contrast, only in-domain PM speech data was used for Spanish. Additionally, it should be noted that the speakers related to the AM data are different from those selected to evaluate the CM models. The statistics of the AM train data are summarised in Table 1.

On the other hand, the *language model* (LM) consisted of 5-gram models computed with the SRILM toolkit ([40]). It is worth mentioning that a common LM was used for all the lectures of the VL task. However, a different LM was used for the PL task depending on the speaker who delivered the speech. Each different LM was adapted to the speaker by exploiting the textual content in the slides available for these PM lectures [9].

The evaluation of CMs has been carried out over a distinct corpus from the data used to build the ASR systems. This corpus was split into *training*, *development* and *test* partitions in a balanced way for each of the speakers (statistics are summarised in Table 2). As a measure of the difficulty of the task, it should be noted that about 25% of the words of each test are not found in the training sets . The *word error rates* (WER) on the automatic transcripts of the VL and PL test sets were 29.97% and 11.83%, respectively.

4.2. Evaluation of CMs

For the purpose of evaluation, the recognised words must be labelled as correct or incorrect. The labelling was computed as the tagging error on the automatic transcripts compared to the reference transcripts based on the optimal Levenshtein alignments. Additionally, class prediction (correct, $c = 1$, or incorrect, $c = 0$) is carried out by minimising the Bayes risk as follows:

$$c^* = \begin{cases} \text{correct} & \mathsf{if}\ \ p(c = 1|w, \mathbf{x}, s) > \tau \\ \text{incorrect} & \text{ow} \end{cases} \tag{16}$$

Note that the speaker dependence in Eq. (16) is not present in the case of speaker independent models. The threshold τ can be empirically estimated

Table 1: *Acoustic data statistics for the English and Spanish ASR systems.*

	videoLectures.net				poliMedia			
Set	Spks	Dur.	Words	Voc.	Spks	Dur.	Words	Voc.
ASR data	4034	427h	2.8M	41K	73	107h	936K	27K

Table 2: *Data partitions for the VL and PL evaluation tasks.*

	videoLectures.net				poliMedia			
Set	Spks	Dur.	Words	Voc.	Spks	Dur.	Words	Voc.
Train	8	3.9h	34K	4K	29	20h	117K	13K
Dev.	8	1.3h	11K	2K	29	6.5h	59K	6K
Test	8	1.3h	11K	2K	29	6.7h	59K	6K

on the development set. However, this was only necessary for the generative models, because the optimal threshold for the discriminative models (LR and LR+spk) resulted always very close to 0.5 due to the MSE training criterion.

The performance of CMs has been tested based on the following evaluation metrics:

- Classification error rate *(CER)*: The relative number of wrongly classified samples on an evaluation sample set, given the rule in (16). It is the direct natural metric to assess the performance of two classifiers: the higher the value, the worse. A simple way to estimate the goodness of a classifier is to compare the CER value to the *relative number of incorrect samples* produced by the system (usually referred as the "baseline"). Unfortunately, the CER as a metric has some flaws: results cannot be directly compared for different tests sets; and the CER is very sensitive to the test set itself, not only to the classifier.

- Area under the ROC curve *(AROC)*: The area under the *Receiving Operating Characteristic* (ROC) curve [41]. Briefly, the ROC curve is the set of points in the *False Positive Rate (FPR)-True Positive Rate (TPR)* space, yielded by the classification for every possible different value of the classification threshold τ. The AROC is usually normalised within $[0, 100]$, 100 being a perfect classification and 50 a random classification. The AROC has been a commonly used metric to evaluate the replicability of the CER results. Nonetheless, this metric has been severely criticised since it can give potentially misleading results if ROC curves cross, and it is incoherent in terms of misclassification costs [42].

- h-measure [42]: Normalised metric which is proportional to the overall misclassification loss incurred when using an optimal threshold (which depends on the costs) averaged by a certain function $u(c)$ over the cost ratio $c \in [0, 1]$, $c = c_0/(c_0 + c_1)$ and (c_0, c_1) being the misclassification costs. For the common case in which it cannot be derived which kind of misclassifications are preferable (false positive, or false negatives, etc.), the author proposes a normalised symmetric function $u(c) \propto \beta(c; 2, 2) \propto (c-c^2)$. This measure was proposed to avoid the issue of the AROC metric, since it is proportional to the expectation of the overall misclassification loss weighted by a function depending on the distribution of the scores.

Thus, the weight function to measure the AROC depends on the classifier to be tested.

- Normalised cross entropy *(NCE)* [43]: Metric proportional to the cross entropy of the classified set. This metric is related to the average log distance of the score to the true class. NCE equals 1 for a perfect classification in which the predicted posteriors of the correct class score 1 for the correct samples and 0 for the incorrect. Unfortunately, the lowest value is unbounded, since it involves the sum of the logarithm of zero or arbitrarily low values for samples which scored high on the opposite class to the true one. Despite this flaw (noticed shortly after its publication [44]), it is still widely used.

4.3. Results

Experiments have been carried out computing the set of input scores that performed the best for the NB model in [18].

- SP: Word acoustic log-score per time frame (10-ms).
- D: Duration (in ms.) of the word per phone.
- NL: Length of the N-gram in which the word has been decoded.
- PAvg: Word posterior probability computed as the average of frame-based posteriors [10].
- PMax: Like PAvg but using the maximum instead of the average [10].

The NL score is not exactly the same as that used in [18], since the length of the N-gram is used instead of the Boolean feature representing the LM back-off behaviour.

Table 3 summarises the performance of the proposed models on the VL and PL test sets in terms of the different metrics presented in Section 4.2. We also include results from additional experiments using *Conditional Random Field* (CRF) models which, as stated in recent publications, are of particular importance [23, 24, 25, 45] [1]. It is worth noting that all models have been compared under identical conditions. To assess statistical significance of results, 95% confidence intervals are included for the CER% evaluation metric.

From the results in Table 3, it can be stated that speaker-adapted models outperform their non-adapted counterparts. This is true, indeed, for all models and all evaluation metrics, and also holds for both, VL and PL tasks. Statistically speaking, this statement is significant to a great extent, especially in the case of VL. In this case, in terms of CER%, the best results are: 14.99, with CRF+spk, and 14.82 with LR+spk. These figures are clearly below the lower limit of the 95% confidence intervals for CRF and LR, respectively. On the other hand, the results on PL are similar, though the CRF+spk result overlap

[1] Both, CRF++ and wapiti toolkits were tested. Results presented here correspond to wapiti toolkit (https://wapiti.limsi.fr/), which in turn outperformed CRF++. The optimisation algorithm used was RPROP+ too with L2 regularisation. The optimisation criterion was Maximum log-Likelihood conditional Estimate.

Table 3: *Performance of the models on VL and PL tasks.*

TASK	MODEL	CER%	CER% 95%CI	AROC%	h	NCE
VL	NB	17.27	$[16.57, 17.97]$	85.4	0.37	0.17
	CRF	16.62	$[15.93, 17.31]$	86.2	0.39	0.31
	LR	16.43	$[15.75, 17.11]$	86.4	0.40	0.32
	NB+spk	16.56	$[15.87, 17.25]$	86.2	0.39	0.19
	CRF+spk	14.99	$[14.33, 15.65]$	88.1	0.44	0.36
	LR+spk	14.82	$[14.16, 15.48]$	88.2	0.45	0.36
PL	NB	8.14	$[\ 7.92, 8.36]$	84.9	0.30	0.07
	CRF	7.99	$[\ 7.77, 8.21]$	85.9	0.31	0.29
	LR	7.89	$[\ 7.67, 8.11]$	85.5	0.31	0.29
	NB+spk	8.09	$[\ 7.87, 8.31]$	85.7	0.31	0.10
	CRF+spk	7.97	$[\ 7.75, 8.19]$	86.9	0.33	0.30
	LR+spk	7.81	$[\ 7.59, 8.03]$	86.4	0.32	0.30

the CER% confidence interval for CRF at its lower half, and the same happens with LR+spk. This might be influenced by the comparatively low values of CER% on PL for all models.

Another conclusion that can be drawn from Table 3 is that the NB model is clearly superseded by CRF and LR, and that this also holds for their speaker-adapted versions. Given that the LR model is designed as a discriminatively trained version of NB, this result was well expected. On the other hand, although LR(+spk) results are slightly but consistently better than those of CRF(+spk), there is no clear statistical evidence to support its superiority. Indeed, the main difference between them is the training criterion used which, from our experiments, has little effect on the results.

The ROC curves of the NB(+spk), CRF(+spk) and LR(+spk) models are depicted in Fig. 1 and Fig. 2 for VL and PL, respectively. The classification thresholds adjusted from the development data (operating points) and the optimal ones are also plotted. As can be observed, the speaker-adapted models show better performance than their basic, non-adapted counterparts for nearly all possible classification thresholds.

Table 4 shows detailed results on the VL test, at speaker level, using the CER evaluation metric. As above, the best results are achieved by LR+spk and CRF+spk. The results at speaker level using other evaluation metrics are similar and are omitted for simplicity.

Table 4: *CER in [%] for each speaker on the VL test set.*

Speaker	1	2	3	4	5	6	7	8
Baseline	15.37	12.79	21.94	16.10	48.76	22.72	31.86	45.89
NB	14.01	12.72	15.91	13.10	27.24	16.15	22.43	30.78
CRF	13.37	12.37	16.20	13.88	25.11	15.96	19.71	28.49
LR	13.00	12.23	16.20	13.55	24.68	15.33	20.31	29.06
NB+spk	13.91	11.74	15.56	13.21	22.03	16.15	21.84	25.62
CRF+spk	12.64	11.81	14.41	12.83	19.56	15.52	18.86	21.99
LR+spk	12.73	11.39	14.41	12.38	19.73	14.77	18.61	23.14

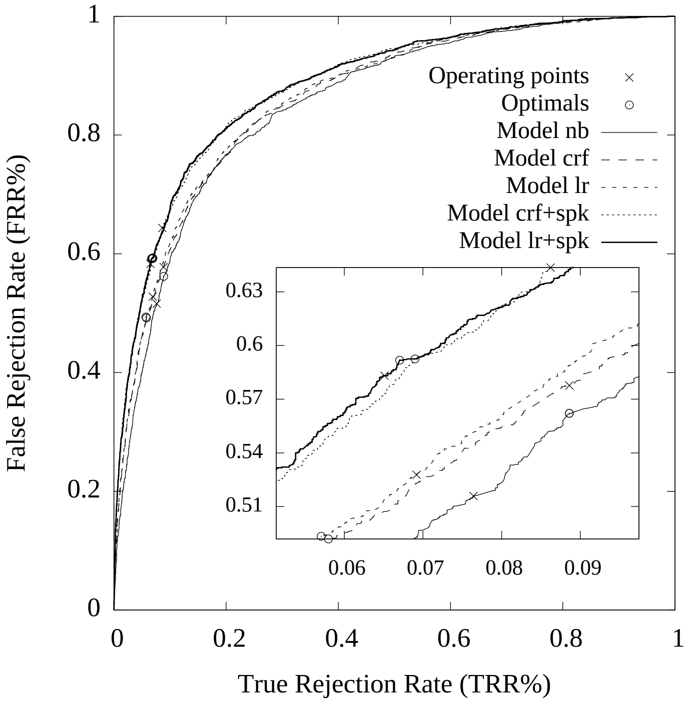

Figure 1: *ROC curves on the videoLectures.net test set.*

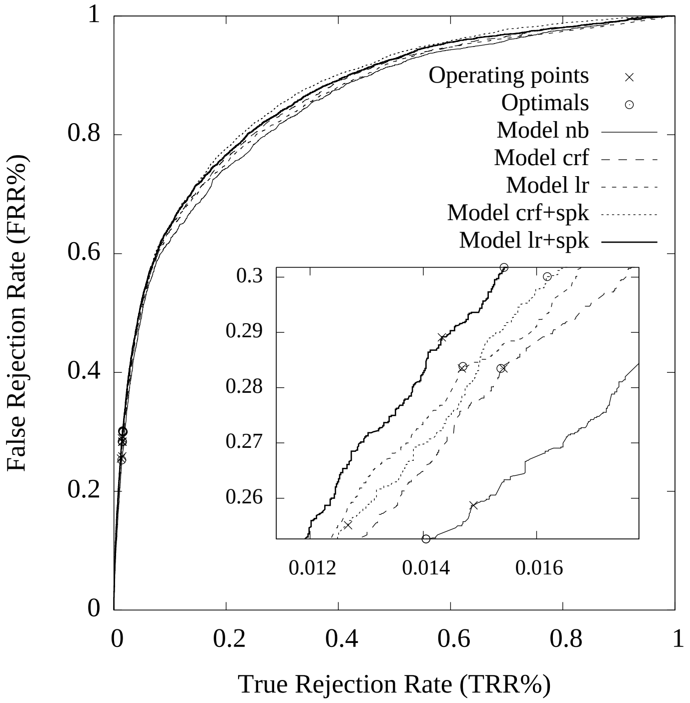

Figure 2: *ROC curves on the poliMedia test set.*

5. Interactive Speech Transcription Application

With the aim of measuring the benefits of the LR+spk model in a practical application, we have evaluated its performance in an *interactive speech transcription* (IST) setting applied within the EU project transLectures. In this setting, users devote a limited amount of effort to supervising a given percentage of words of the automatic transcriptions. User effort is optimised by ordering the speech segments selected for supervision from lower to higher reliability based on CMs.

The VL test set has been used for the assessment of the NB and LR+spk models. Corrections were performed by means of a simulated user in a similar way to [46].

The final quality (measured in WER) of partially supervised transcriptions resulting for different percentages of supervised words is depicted on Fig. 3. The figure assesses the behaviour when using the NB, LR+spk or CRF+spk (wapiti+spk) models to compute CMs. A random strategy corresponding to a sequential supervision of the words is also depicted.

From Fig. 3, it can be stated that the LR+spk and CRF+spk models perform similarly, and that they both outperform the NB model for any level of user effort (percentage of supervised words). In particular, for the reasonable range of percentages from 10% to 20%, the LR+spk and CRF+spk produce relative WER improvements between 2% and 7%.

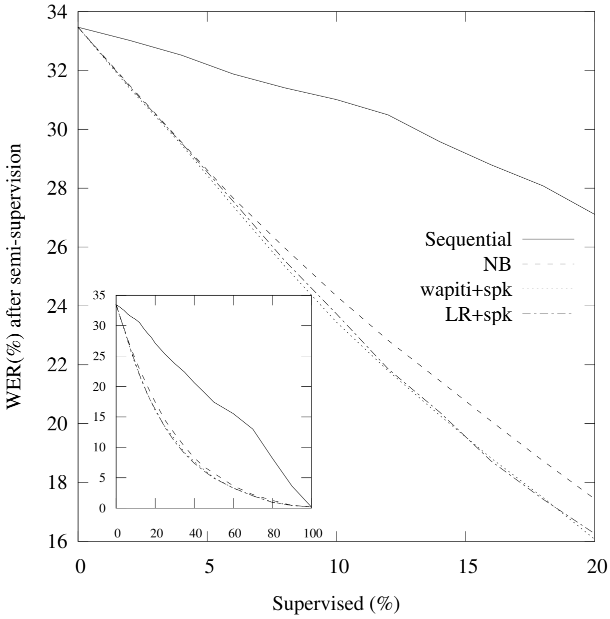

Figure 3: *Resulting WER for the partial supervision of the VL test set in a reasonable range of effort. The sub-figure shows the behaviour under all percentages up to full supervision.*

6. Conclusions

We have introduced a new particular logistic regression model to improve the reliability of the confidence measures for automatic speech recognition. Also, as a main contribution, we have proposed the use of speaker-adapted models.

The experiments have shown that speaker-adapted models outperform their non-adapted counterparts on two difficult tasks from English (videoLectures.net) and Spanish (poliMedia) educational lectures. The proposed logistic regression model achieved comparatively good results.

Finally, a simple real application of interactive speech transcription guided by confidence measures has confirmed that the gains obtained by the proposed models translate into a noticeable improvement of the resulting semi-supervised transcriptions for an equal level of user effort.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 287755. Also supported by the Spanish MINECO (iTrans2 TIN2009-14511 and Active2Trans TIN2012-31723) research projects and the FPI Scholarship BES-2010-033005.

References

[1] A. Rousseau, Lium's systems for the iwslt 2011 speech translation tasks, in: International Workshop on Spoken Language Translation, San Francisco (USA), 2011.

[2] M. Sundermeyer, M. Nußbaum-Thom, S. Wiesler, C. Plahl, A. Mousa, S. Hahn, D. Nolden, R. Schlüter, H. Ney, The RWTH 2010 quaero ASR evaluation system for English, French, and German, Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Prague, Czech Republic (2011) 2212–2215.

[3] P. Swietojanski, A. Ghoshal, S. Renals, Revisiting hybrid and GMM-HMM system combination techniques, in: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, IEEE, 2013, pp. 6744–6748.

[4] C. Leggetter, P. Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models, Computer speech and language 9 (2) (1995) 171.

[5] M. Gales, Maximum likelihood linear transformations for hmm-based speech recognition, Computer Speech and Language 12 (1998) 75–98.

[6] J. Gauvain, C. Lee, Maximum a posteriori estimation of multivariate gaussian mixture observations of markov chains, IEEE Trans. on Speech and Audio Processing 2 (2) (1994) 291–298.

[7] V. Digalakis, D. Rtischev, L. Neumeyer, Speaker adaptation using constrained reestimation of gaussian mixtures, IEEE Trans. on Speech and Audio Processing 3 (5) (1995) 357–366.

[8] S. Wiesler, K. Irie, Z. Tüske, R. Schlüter, H. Ney, The rwth english lecture recognition system, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, Florence, Italy, 2014, pp. 3322–3326.

[9] A. Martinez-Villaronga, M. del Agua, J. Andres-Ferrer, A. Juan, Language model adaptation for video lectures transcription, in: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, 2013, pp. 8450–8454. `doi:10.1109/ICASSP.2013.6639314`.

[10] F. Wessel, R. Schluter, K. Macherey, H. Ney, Confidence measures for large vocabulary continuous speech recognition, Speech and Audio Processing, IEEE Transactions on 9 (3) (2001) 288–298. `doi:10.1109/89.906002`.

[11] T.-Y. Kim, H. Ko, Bayesian fusion of confidence measures for speech recognition, IEEE Signal Processing Letters 12 (2005) 871–874.

[12] J. Gao, Q. Zhao, R. Xu, Y. Yan, Improved lattice-based confidence measure for speech recognition via a lattice cutoff procedure, in: Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on, Vol. 4, 2009, pp. 473–476. `doi:10.1109/FSKD.2009.367`.

[13] M. Bardideh, F. Razzazi, H. Ghassemian, An svm based confidence measure for continuous speech recognition, in: Signal Processing and Communications, 2007. ICSPC 2007. IEEE International Conference on, 2007, pp. 1015–1018. `doi:10.1109/ICSPC.2007.4728494`.

[14] Z.-G. Wang, C. Liu, H.-K. Wang, Y. Hu, L.-R. Dai, Phonetic clustering based confidence measure for embedded speech recognition, Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on (2010) 186–189.

[15] Z. Junfeng, Z. Yeping, A multi-confidence feature combination rejection method for robust speech recognition, in: Transportation, Mechanical, and Electrical Engineering (TMEE), 2011 International Conference on, 2011, pp. 2556–2559. `doi:10.1109/TMEE.2011.6199743`.

[16] K. A. Yadav, M. Patil, Confidence calibration measures to improve speech recognition, in: Communications and Signal Processing (ICCSP), 2013 International Conference on, IEEE, 2013, pp. 826–829.

[17] H. Jiang, Confidence measures for speech recognition: A survey, Speech communication 45 (4) (2005) 455–470.

[18] A. Sanchis, A. Juan, E. Vidal, A Word-Based Naïve Bayes Classifier for Confidence Estimation in Speech Recognition, Audio, Speech, and Language Processing, IEEE Transactions on 20 (2) (2012) 565–574.

[19] G. Heigold, P. Dreuw, S. Hahn, R. Schüter, H. Ney, Margin-Based Discriminative Training for String Recognition (2010) 1–10.

[20] S. Guiasu, A. Shenitzer, The principle of maximum entropy, The Mathematical Intelligencer 7 (1) (1985) 42–48. doi:10.1007/BF03023004.
URL http://dx.doi.org/10.1007/BF03023004

[21] D. Yu, J. Li, L. Deng, Calibration of confidence measures in speech recognition, Audio, Speech, and Language Processing, IEEE Transactions on 19 (8) (2011) 2461–2473.

[22] J. A. Silvestre-Cerda, A. Perez, M. Jimenez, C. Turro, A. Juan, J. Civera, A System Architecture to Support Cost-Effective Transcription and Translation of Large Video Lecture Repositories, in: Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on, IEEE, 2013, pp. 3994–3999.

[23] M. S. Seigel, Confidence estimation for automatic speech recognition hypotheses, Ph.D. thesis, Department of Engineering, University of Cambridge (2013).

[24] M. S. Seigel, P. C. Woodland, Combining information sources for confidence estimation with CRF models., in: INTERSPEECH, 2011, pp. 905–908.

[25] J. Fayolle, F. Moreau, C. Raymond, G. Gravier, P. Gros, CRF-based combination of contextual features to improve a posteriori word-level confidence measures., in: INTERSPEECH, 2010, pp. 1942–1945.

[26] A. Sanchis, A. Juan, E. Vidal, Estimation of confidence measures for machine translation, in: Proceedings of the MT Summit XI, 2007.

[27] C. Estienne, A. Sanchis, A. Juan, E. Vidaf, Maximum entropy models for speech confidence estimation, Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on (2008) 4421–4424.

[28] C. Igel, M. Hüsken, Empirical evaluation of the improved RPROP learning algorithms, Neurocomputing 50 (2003) 105–123.

[29] J. N. Darroch, D. Ratcliff, Generalized Iterative Scaling for Log-Linear Models, The Annals of Mathematical Statistics 43 (5) (1972) 1470–1480. doi:10.1214/aoms/1177692379.
URL http://projecteuclid.org/euclid.aoms/1177692379

[30] S. Wiesler, A. Richard, R. Schluter, H. Ney, A critical evaluation of stochastic algorithms for convex optimization, in: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, IEEE, 2013, pp. 6955–6959.

[31] J. A. Silvestre-Cerdà, M. A. del Agua, G. Garcés, G. Gascó, A. Giménez, A. Martínez, A. Pérez, I. Sánchez, N. Serrano, R. Spencer, J. D. Valor, J. Andrés-Ferrer, J. Civera, A. Sanchis, A. Juan, translectures, in: Online Proc. of Advances in Speech and Language Technologies for Iberian Languages (IBERSPEECH 2012), Madrid (Spain), 2012, pp. 345–351.
URL http://iberspeech2012.ii.uam.es/IberSPEECH2012_OnlineProceedings.pdf

[32] T. T.-U. team. The translectures upv toolkit (tlk) [online].

[33] G. E. Dahl, S. Member, D. Yu, S. Member, L. Deng, A. Acero, Context-dependent pre-trained deep neural networks for large vocabulary speech recognition, in: IEEE Transactions on Audio, Speech, and Language Processing, 2012.

[34] G. Stemmer, F. Brugnara, D. Giuliani, Adaptive training using simple target models, in: Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on, Vol. 1, 2005, pp. 997 – 1000.

[35] D. Giuliani, M. Gerosa, F. Brugnara, Speaker normalization through constrained MLLR based transforms., in: INTERSPEECH, 2004.

[36] A. Rousseau, P. Deléglise, Y. Estève, TED-LIUM: an Automatic Speech Recognition dedicated corpus., LREC.

[37] B. Ramabhadran, O. Siohan, A. Sethy, The IBM 2007 speech transcription system for European parliamentary speeches, in: IEEE Workshop on ASRU, 2007, pp. 472–477.

[38] TC-STAR Evaluation Report, `www.tcstar.org/documents/D30.pdf`.

[39] Voxforge, `www.voxforge.org/`.

[40] A. Stolcke, SRILM – an extensible language modeling toolkit, in: Proceedings of ICSLP, Vol. 2, Denver, USA, 2002, pp. 901–904.

[41] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (8) (2006) 861–874.

[42] D. Hand, Measuring classifier performance: a coherent alternative to the area under the roc curve, Machine Learning 77 (1) (2009) 103–123. `doi:10.1007/s10994-009-5119-5`.
URL `http://dx.doi.org/10.1007/s10994-009-5119-5`

[43] NCE, `www.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm`.

[44] F. Wessel, K. Macherey, R. Schluter, Using word probabilities as confidence measures, in: Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, Vol. 1, IEEE, 1998, pp. 225–228.

[45] T. Lavergne, O. Cappé, F. Yvon, Practical very large scale CRFs, in: Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, 2010, pp. 504–513.
URL `http://www.aclweb.org/anthology/P10-1052`

[46] I. Sanchez-Cortina, N. Serrano, A. Sanchis, A. Juan, A prototype for interactive speech transcription balancing error and supervision effort, in: IUI '12: Proceedings of the 2012 ACM international conference on Intelligent User Interfaces, ACM, 2012. `doi:10.1145/2166966.2167035`.
URL `http://dl.acm.org/citation.cfm?id=2167035`

Vitae

Isaias Sanchez-Cortina is a Ph.D. candidate at the Universitat Politècnica de València (UPV). He received his Physics Degree from Universitat de València (2007) and was a researcher on optics of the human visual system. After that, he worked on machine translation and speech recognition research and obtained his M.Sc. in Artificial Intelligence, Pattern Recognition and Digital Imaging (UPV, 2012). His current research interests include data analysis, machine learning and pattern recognition.

Jesús Andrés-Ferrer received his Degree in Computer Science (2004), his M.Sc. in Artificial Intelligence, Pattern Recognition and Digital Imaging (2008) and his European Ph.D. degree in Computer Science (2010, with an award in 2011 by AERFAI) from the Universitat Politècnica de València (UPV). He is coauthor of many articles and he has been actively involved in several projects (erudito.com, transLectures, etc.). Currently, he is working at Nuance Communications. His research interests include machine learning and pattern recognition techniques applied to statistical machine translation, speech recognition, handwritten recognition and, especially, language modelling.

Alberto Sanchis is a Ph.D. Associate Professor of Computer Science at the Universitat Politècnica de València (Spain). He obtained his PhD in 2004, and he is member of the Machine Learning and Language Processing (MLLP) group. He is the co-author of more than 50 articles in international journals and conferences. He has participated in 5 European projects and more than 10 Spanish projects. He is now leading a project funded by the Spanish Government on active interaction for speech transcription and translation.

Alfons Juan is a Professor of Computer Science at the Universitat Politècnica de València (UPV), where he obtained his PhD in 2000 and leads a group of about 14 researchers on Machine Learning and Language Processing (MLLP). He has participated in more than 30 research projects and has published over 130 articles in international journals and conferences. He has also been an advisor for 8 PhD thesis on different MLLP topics. His most recent work includes the leadership of the València node of the EU Network of Excellence PASCAL2, as well as the coordination of the EU research project transLectures.