Automatic Quality Estimation for ASR System Combination

Shahab Jalalvand[a,b,*], Matteo Negri[a,*], Daniele Falavigna[a], Marco Matassoni[a], Marco Turchi[a]

[a]*FBK-Fondazione Bruno Kessler, Trento, Italy*
[b]*University of Trento, Italy*

arXiv:1706.07238v1 [cs.CL] 22 Jun 2017

Abstract

Recognizer Output Voting Error Reduction (ROVER) has been widely used for system combination in automatic speech recognition (ASR). In order to select the most appropriate words to insert at each position in the output transcriptions, some ROVER extensions rely on critical information such as confidence scores and other ASR decoder features. This information, which is not always available, highly depends on the decoding process and sometimes tends to overestimate the real quality of the recognized words. In this paper we propose a novel variant of ROVER that takes advantage of ASR quality estimation (QE) for ranking the transcriptions at "segment level" instead of: *i)* relying on confidence scores, or *ii)* feeding ROVER with randomly ordered hypotheses. We first introduce an effective set of features to compensate for the absence of ASR decoder information. Then, we apply QE techniques to perform accurate hypothesis ranking at segment-level before starting the fusion process. The evaluation is carried out on two different tasks, in which we respectively combine hypotheses coming from independent ASR systems and multi-microphone recordings. In both tasks, it is assumed that the ASR decoder information is not available. The proposed approach significantly outperforms standard ROVER and it is competitive with two strong oracles that exploit prior knowledge about the real quality of the hypotheses to be combined. Compared to standard ROVER, the absolute WER improvements in the two evaluation scenarios range from 0.5% to 7.3%.

Keywords: automatic speech recognition, quality estimation, system combination

1. Introduction

The application of ASR systems in our daily life is steadily increasing. Voice search engines, voice question answering, broadcast news transcriptions, video/TV programs subtitling, meeting transcriptions and spoken dialog systems are just some of the many applications involving ASR technology. In such applications, the quality of transcriptions and

---

the availability of fast and accurate automatic evaluation methods represent two crucial interconnected needs. Automatic ASR evaluation, indeed, does not only represent a way to assess system performance *a posteriori*, but also a way to improve it. For instance, automatic assessment methods can be used to select audio data for unsupervised training [43, 25] and active learning of acoustic models [58, 24] or, as in the case of this paper, to combine multiple transcription hypotheses into a single and more accurate one.

In order to synthetically obtain more accurate transcriptions, ASR systems diversity and complementarity have been exploited in different ways [2]. The combination of multiple hypotheses coming from independent sources usually leads to significant improvement compared to the output of each individual system. ROVER, the most popular ASR system combination approach, performs hypothesis fusion by first building a word confusion network (CN) from the *1*-best hypotheses of the ASR systems entering the combination and then by selecting the best word in each CN bin via majority voting [27]. Word confidence scores, when available, can be exploited to inform the process by setting up a weighted majority voting scheme. Although several enhancements of this general strategy have been successfully proposed, ROVER-based hypothesis combination methods still suffer from some limitations that this paper aims to overcome.

The first one depends on how they are generally implemented: the hypothesis combination process considers the first input candidate as a "skeleton" to align the other hypotheses in a greedy manner. For this reason, depending on the order in which the hypotheses are considered when feeding the algorithm, the resulting combination can show large variations in quality. This raises the need of reliable techniques for ranking the input hypotheses before starting the fusion process.

The second limitation relates to the granularity of the input fed into ROVER: in the standard setting, transcription hypotheses correspond to entire audio recordings, whose duration can span up to hours. However, when manipulating long utterance transcriptions, the skeleton used to initialise the process can feature significant local variations in terms of quality. As a consequence, it may happen that the worst hypothesis for an entire audio recording (*i.e.* globally) is the best one for one or more passages (*i.e.* locally). This raises the need of strategies for operating at higher levels of granularity (*e.g.* segments spanning over few seconds) in order to take full advantage of local quality differences between the input hypotheses.

The third limitation concerns the applicability of ROVER-based hypothesis combination techniques: results' quality significantly increases when the availability of confidence scores makes it possible to set up weighted majority voting schemes. However, having access to the ASR systems' inner workings is a rigid constraint that limits the applicability of hypothesis fusion to scenarios in which the input transcriptions are produced by known ASR tools. Often, however, the hypotheses come from "black-box" systems, without additional scores.[1] This raises the need of methods that are independent from confidence information, but still

---

[1]For instance, this is the case of the steadily increasing volume of Youtube videos for which automatic captions are provided by black-box ASR technology. In 2012, more than 157 million videos were already accessible with auto-captions in 10 languages (source: http://goo.gl/5Wlkjl).

capable to achieve good results with simple frequency-based voting.

Finally, it is worth noting that the confidence scores proposed by previous ASR literature [22, 73], even when applicable, only indicate how confident the system is about its own output. This can be a biased perspective (influenced by individual decoder features), producing scores that are not comparable across different systems. External and system-independent measures of goodness would represent a more reliable alternative when comparable and objective ASR quality judgements are required.

To cope with these issues, in [53] we proposed a reference-free and confidence-independent ASR quality estimation (QE) method, in which a supervised regression model is used to predict the word error rate (WER) of automatically transcribed audio recordings. Experimental results in different evaluation settings showed that our QE predictions can closely approximate the true WER scores calculated over reference transcripts. Building on these positive results, in [39] we applied ASR QE to inform system combination with ROVER, outlining the framework that this paper aims to extend and refine. Our study focused on comparing the standard *system-level* ROVER with an alternative *segment-level* strategy that uses ASR QE to rank the input hypotheses before starting the fusion process. Our method was applied to combine the transcriptions of English TED talks produced by the eight participants in the IWSLT2013 evaluation campaign.[2] Its results outperformed the standard ROVER (based on averaging the results over a large number of random rankings of ASR outputs) and significantly approached oracle upper-bounds. A preliminary discussion about the influence of hypothesis diversity on the fusion results was also proposed (an issue theoretically studied also in [2]). However, a deeper investigation on how to better exploit the trade-off between the number of combined hypotheses and their diversity was left for future works.

To further assess the effectiveness and the general applicability of the proposed QE-informed ROVER, this paper extends our experiments to a new and different scenario involving multiple distant microphone (MDM) speech recognition in noisy environments. The new experiments are carried out with the data delivered for the 3rd CHiME challenge[3] [4], in which six microphones, placed on a tablet PC, were used to record sentences of the Wall Street Journal (WSJ) corpus [54] uttered by different speakers in four environments (bus, cafe, pedestrian area and street junction). In this case, the diversity among the combined ASR hypotheses is caused by different positions of the microphones that acquire the audio signals, while the ASR system that processes the different audio streams remains unchanged.

Although MDM hypothesis combination can be considered as an alternative to the signal enhancement techniques proposed in the past [76, 42, 50], in this work we show that QE-informed ROVER yields further improvements, even after including the enhanced channels in the combination. Similar to [39], we first use ASR QE techniques to predict the quality of each channel transcription (including the enhanced channels) and then we rank them before applying ROVER. To the best of our knowledge, this is the first time that ASR QE

---

[2]The International Workshop on Spoken Language Translation (IWSLT – `http://workshop2013.iwslt.org/`) is a yearly workshop associated with an open evaluation campaign on spoken language translation.

[3]The CHiME Speech Separation and Recognition Challenges are international initiatives proposed in 2011, 2013, 2015 and 2016 – `http://spandh.dcs.shef.ac.uk/chime_challenge/`.

is applied for ranking MDM hypotheses.

Building on the positive results presented in [39] and by discussing a new set of experiments, this paper shows that:

- ASR QE can support system combination with ROVER by providing reliable ranks of the hypotheses produced by different ASR systems. For the automatic transcriptions of TED talks in the IWSLT dataset, the absolute WER reductions with respect to the best individual system and standard random ROVER are respectively 1.6% and 0.5%;

- ASR QE consistently works well also for hypothesis-level combination in the MDM speech recognition scenario (same ASR system but multiple microphones). On CHiME-3 data, The WER reductions with respect to the best enhanced channel and random ROVER are respectively 1.7% and 7.3%.

Moreover, by extending our analysis to new problems, in this work we overcome two limitations of method originally proposed in [39]:

- Dealing with tied ranks. Often, in the training data, several transcriptions of the same segment have the same WER in spite of their different quality. In such cases, having the same label associated to different features reduces the performance of our ranking algorithms. We propose a simple solution to this problem, which allows us to obtain better and coherent results over the two datasets.

- Finding an optimum level of combination. Usually, combining all the available hypotheses does not lead to the best solution. ROVER, indeed, benefits not only from the quality of the combined transcriptions, but also from their diversity. By investigating this trade-off, we provide an empirical method to select an optimum number of hypotheses to combine for each segment (*i.e.* locally). Our results closely approximate those obtained, on both datasets, by the best combinations at global level.

The paper is organized as follows. After an overview of related work in §2, we shortly introduce how ROVER works in §3. This allows us to set the stage for formalizing our approach to system combination and describe its overall workflow in §4. Then, together with numeric evidence emphasizing the potential of our QE-informed approach, in §5 we provide more details about the algorithms and the features used. Our experimental setting is fully described in §6, followed by a discussion of the results achieved on IWSLT and CHiME-3 data in §7. Finally, §8 and §9 respectively investigate new problems (dealing with tied ranks and automatically finding the optimum level for combination) and conclude the paper with final remarks.

2. Related work

Scientific literature related to the work described in this article spans over several lines of investigation such as ASR quality estimation, ASR system combination, signal enhancement and speech recognition with multiple distant microphones in noisy environments.

Quality Estimation. Differently from confidence estimation, which has been widely studied over the last two decades [22, 73], ASR QE is a rather new task. The problem, at its core, shows a strong parallelism with QE in the machine translation (MT) field, where the goal of bypassing the need of manually-created reference translations has motivated a large body of research. The motivations (cost effective quality prediction at run-time) and the methods (supervised learning, either as regression or multiclass classification) are indeed the same in both fields. For a complete overview of the current approaches to MT QE, we refer the reader to the comprehensive overviews published within the yearly Workshops on Statistical Machine Translation [16, 9, 10, 11] and to the works dealing with quality prediction at word level [68, 3], sentence level [64, 67, 61, 62] and document level [60].

In the ASR field, confidence estimation strongly relies on knowledge about the inner workings of the decoder that produces the transcriptions. Such knowledge includes acoustic and language model scores and word timing information associated to arcs of word lattices. QE, instead, approaches the problem as a system-independent task, in which this information is not necessarily accessible. In the first work along this direction [53], we explored ASR QE as a supervised regression problem in which the WER of an utterance transcription has to be automatically predicted. Given a set of (*signal*, *transcription*, *WER*) triplets as training instances, different algorithms were evaluated on test data consisting of unseen (*signal*, *transcription*) pairs. All models used an effective set of signal/textual/hybrid features, named *black-box* features in contrast with the *glass-box* ones that model the internal behavior of the decoder. Results showed that ASR QE predictions can closely approximate the "true" WER scores calculated over reference transcripts, outperforming a strong baseline in a variety of test conditions of increasing complexity. Interestingly, when combined with the glass-box features (*i.e.* confidence scores), the predictions are better than those obtained by using confidence scores alone.

In [63], ASR QE was explored by focusing on the problem of domain mismatches between training and test data. Indeed, as pointed out by [53], simple supervised learning methods are very sensitive to large variations in the distribution of the instances in the two sets (both at the level of labels and at the level of features). The proposed solution relies on multitask learning to train robust models that exploit similarities and differences between possibly related tasks, transferring knowledge across them. Results show that the approach is able to take advantage of data coming from such heterogeneous domains and to significantly improve over single-task learning baselines both in regression and in classification. These findings indicate the reliability of ASR QE in particularly challenging test conditions that are out of the scope of this work. Casting ASR QE for system combination as a multitask learning problem, however, is certainly an aspect that we will explore in the future.

System combination. Most of the approaches proposed in the past for combining multiple ASR outputs make use of word lattices [44, 12]. The underlying idea is to merge the word lattices generated by different ASR systems into a single one, which is then traversed to search for the best path. As an alternative, frame-based system combination [35] tries to minimize a cost function called "time frame word error" (fWER) over a set of word lattices produced by different ASR systems. The method makes it possible to estimate the path

exhibiting the minimum Bayes risk, without the necessity of merging edges of single word lattices. Confusion network combination (CNC) [48, 35, 23] is another widely investigated approach, in which confusion networks built from the individual lattices are aligned instead of single best outputs. More recent approaches [77, 72] use log-linear models, also exploited in machine translation tasks [20], in which the acoustic likelihoods coming from different systems are weighted and summed at frame level. The resulting weighted likelihoods are decoded to produce the final hypotheses. In this way, the combination is applied at a much finer level than that of whole hypotheses (such as CNC or ROVER), resulting in significantly better performance in a standard noise-robust speech recognition task [77]. Note that all combination methods based on the use of word lattices, as well as some extensions of ROVER [78, 34, 1], require to know and have access to the inner structure of ASR decoder. In most cases this is reflected into the computation of confidence measures for the recognized words. Often, however, ASR systems (especially those embedded in commercial applications) do not provide this information. Instead, standard ROVER and the QE-informed ROVER proposed in this paper do not necessarily need to know the ASR inner behaviour.

Another limitation of previous ROVER-based methods is that, assuming a fixed order in the quality of the systems that generated the hypotheses, they do not apply any segment-level ranking, disregarding the possible advantages of operating at a higher granularity level. The work reported in [59] proposes to use the probability given by a LM to solve the ambiguity (see §3) that arises when different words receive the same vote at each confusion bin in the ROVER word network. Inspired by this work, in [39], we presented the first application of ASR QE as a way to inform system combination with ROVER. Different solutions to rank the input hypotheses at *segment level* were compared against the standard ROVER that operates at *system level*. In addition to transferring the QE problem into a machine-learned ranking task, we enriched the feature set used in [53, 63] by adding word-based features previously proposed in [45, 56] for word error detection tasks [32, 66, 37]. The approach, which is summarized in §4, was tested on a specific scenario (the transcription of the English TED talks from IWSLT) and achieved promising results that are also reported in this paper (§7.1). Improving the method, addressing the need for a stopping criterion to avoid entering useless inputs into the ROVER combination and extending the evaluation to new scenarios were left as future works in [39] and represent the main extensions discussed in this paper.

Multiple Distant Microphone (MDM) speech recognition. The use of multiple microphones in ASR is mostly motivated by the necessity to reduce the effects of environmental reverberation and noise over the speech signal to be recognized in distant-talking scenarios. To do this, delay-and-sum beamforming (DS) is a widely used technique to combine multiple recordings. It consists in summing the input signals after having temporarily aligned them in order to compensate for the different times of arrival of the sound waveforms to the various microphones [76]. This approach has been successfully applied to the recognition of voice commands in a domestic environment [69, 52] and it was also used in the CHiME challenge [5]. Another well known approach to generate an "enhanced" speech waveform from multiple recordings is based on minimum variance distortionless response (MVDR) [41]. This solution has been applied as speech enhancement baseline in the CHiME-3 scenario [4].

In addition to DS, MVDR and other signal-based combination approaches [50], the automatic selection of the best channel can be used to perform signal enhancement. The simplest method to do this consists in measuring the signal-to-noise ratio (SNR) of the recorded signals [74], assuming that the channel with the highest SNR will be the easiest to transcribe. Choosing the channel for which the ASR engine has generated the highest confidence score can also represent a viable solution. For instance, in a previous work with data from the aforementioned CHiME-3 challenge, we successfully applied channel selection based on sentence confidence estimation [38] to select the best microphone.

An alternative to signal-level combination and best microphone selection is represented by hypothesis-level combination. Along this direction, [75] proposes to use CNC for lecture transcriptions and [65] extends it with a hybrid approach that leverages beamforming and signal-level diversity to transcribe meetings. In [33], inter-microphone agreement is used to build a confusion network from multiple lattices and improve ASR in a domotic application.

Our QE-informed ROVER solution is also compared against MVDR and DS processing and in addition, we consider both MVDR and DS hypotheses as enhanced channels and we include them in ROVER combination, showing performance improvements with respect to original channel combination.

3. ROVER

Before discussing our QE-informed approach to hypothesis fusion, in this section we introduce ROVER and the way it is commonly used to combine multiple transcriptions. Figure 1 illustrates an example in which three ASR hypotheses (Sys1, Sys2 and Sys3), have to be combined (for the sake of clarity, the words in each hypothesis are represented with letters). Given the input transcriptions ROVER builds a word transition network (WTN) by exploiting iterative dynamic programming. Initially, this is done by selecting the first input hypothesis (Sys1) as a "skeleton" to build an initial base WTN ($WTN0$). Then, the second hypothesis (Sys2) is aligned with $WTN0$ using dynamic programming and Levenshtein distance, resulting in the alignment depicted in Figure 1 ($Align1$), where the asterisks represent both inserted and deleted words. Such alignment is used to build a composite WTN ($WTN1$) by using the symbol "@" to represent insertions and deletions. The process is repeated with the next hypothesis (Sys3), which results in the final composite $WTN2$[4].

The last WTN, resulting from the multiple alignment, can be interpreted as a word confusion network (CN) [48] with a sequence of word bins. Starting from this CN, the combined hypothesis is generated by selecting the best word at each bin via majority voting. In case different words receive the same vote (see the red bin in Figure 1), ROVER gives priority to the word in the first position of the bin (*i.e.* "d" from Sys1). This solution for resolving ties raises one of the important problems addressed in this paper. In fact,

---

[4]Note that to optimally align an hypothesis with a composite WTN a multidimensional dynamic programming would be required, where each dimension is treated as an input hypothesis. Since the implementation of such algorithm is difficult, ROVER [27] uses the approximated solution here summarized.
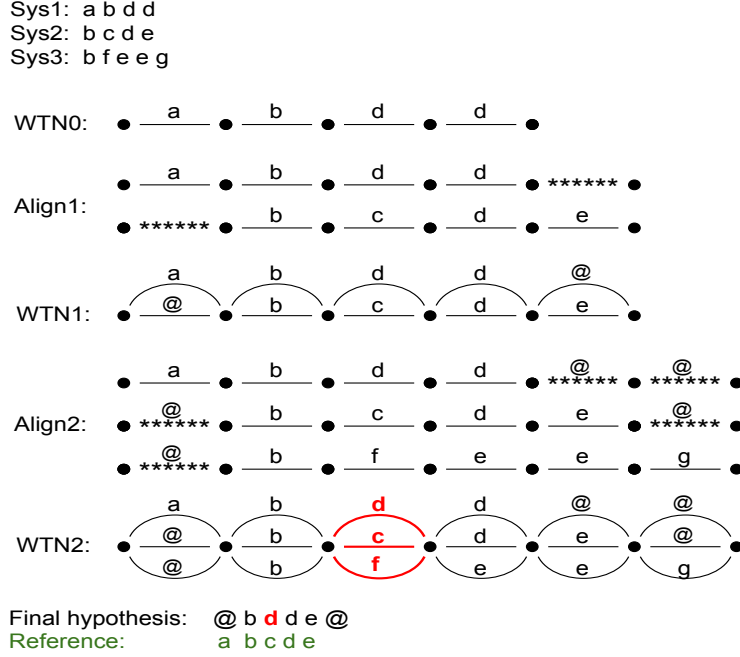
Figure 1: ROVER procedure. $Sys2$ has recognized better than $Sys1$ though it is in the second order. This mistake in the input arrangement, leads to an error in the final hypothesis, while it could be recovered with correct arrangement.

since the first hypothesis used as skeleton highly influences the voting mechanism, the final result will depend on its quality: bad transcriptions can depend from the priority given to a poor-quality skeleton. To mitigate this problem, the approach discussed in the next sections addresses automatic methods for ranking the input hypotheses based on their estimated quality.

Traditionally, ROVER is run by applying random or arbitrary fixed rankings of the input hypotheses at *system-level*. By system-level ranking we refer to an ordering that is assigned to the ASR systems included in the combination. This ranking is the same for all the segments in the dataset. In our example, if the proposed ranking for the three ASR systems is $Sys1 \preceq Sys2 \preceq Sys3$, then this ranking will be kept unchanged for the entire utterance. As a result, ROVER does not exploit local quality differences in the transcriptions of the utterance. In contrast, the approach proposed in the next sections relies on a *segment-level* ranking that dynamically varies from one segment to another based on their predicted quality. This solution makes it possible to take advantage from the fact that different ASR systems (or microphones) show different performance over different portions of the utterance.

4. General workflow

The workflow of our segment-based QE-informed ROVER is illustrated in Figure 2. As previously mentioned we address two application scenarios: *i)* the combination of transcriptions of a single recording coming from N different ASR systems, and *ii)* the combination

of transcriptions of M recordings coming from N different ASR systems. Figure 2 refers to the general case: multiple microphones ($M > 1$), multiple ASR systems ($N > 1$).

Audio recordings are processed by a system that detects low energy regions (applying specific energy thresholds and produce segments of lower duration. Then, each segment is assigned to one of two classes, "speech" or "not speech", and only the speech segments are passed to the ASR systems for transcription.
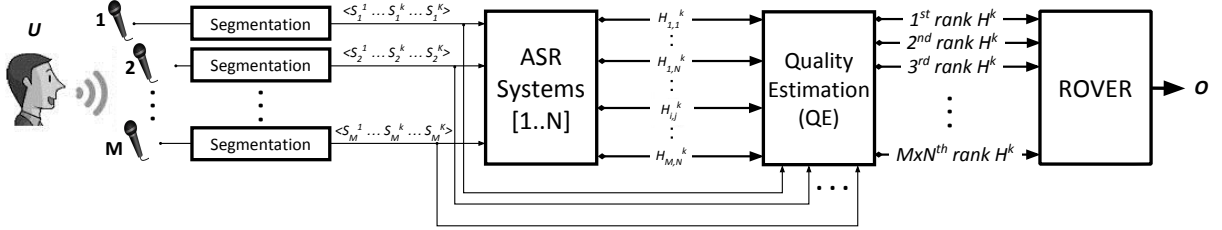

Figure 2: The workflow of segment-based QE-informed ROVER.

The automatically detected speech segments $\{S_1^k \ldots S_M^k\}$ ($1 \leq k \leq K$), resulting from the segmentation of the input utterance U, are sent to N ASR systems producing, for each segment S_j^k, N different hypotheses $\{H_{j,1}^k \ldots H_{j,N}^k\}$.

The signals and the hypotheses are passed to a QE component. The QE model is trained to predict the ranking of the hypotheses by using features extracted from the signals and the transcriptions. Based on the determined ranking, ROVER combines the hypotheses of the kth segment S_j^k ($1 \leq j \leq M$) into o^k. Finally, the combined segment-level hypotheses o^k are concatenated in a single output $O = \{o^1 + o^2.. + o^K\}$ for the entire audio track.

Note that, in order to apply this approach, the same speech segmentation $\{S_1^k \ldots S_M^k\}$ must be shared across all microphones and ASR systems. However, this is not usually true because microphones capture voices at different levels of energy and ASR systems make use of different automatic segmentation algorithms. A possible solution to address this issue is to extract a common segmentation (derived from an enhanced signal or directly available from a manual segmentation).

Instead, when dealing with multiple ASR systems segmentation, an alignment with a reference segmentation needs to be applied. This is the case of the experiments with single microphone multiple ASR systems, as explained in §6.

The general approach depicted in Figure 2 is suitable for a range of conditions in which the hypotheses to be combined come from multiple microphones, multiple ASR systems or any combination of the two. The experimental framework considered in the remainder of this paper focuses on the following two scenarios:

1. Single microphone ($M = 1$), multiple systems ($N > 1$). This is the case of the automatic transcription of TED talks from the IWSLT campaign (see §6.1).
2. Multiple microphones ($M > 1$), multiple systems ($N > 1$). This is the case of the automatic transcription of WSJ articles from the CHiME-3 campaign (see §6.2).

The following sections describe how this general workflow has been instantiated and evaluated in the two scenarios.

5. Segment-level QE-informed ROVER

5.1. Preliminary analysis

To investigate the impact on ROVER results of considering different rankings and different levels of granularity of the input hypotheses, we run some pilot experiments on the same data used in §6, namely: TED talks transcriptions (collected with a close talk microphone) and WSJ read sentences (recorded with distant microphones). In the first case we used the submissions of the IWSLT2013 ASR evaluation campaign [18], where 8 teams submitted their ASR results, each containing hypothesized word sequences and related time boundaries for each talk. In the second case, we used the data provided for the 3rd CHiME challenge [4], where each utterance is recorded by 6 microphones embedded on a tablet PC. One of these microphones, however, is ignored in all our experiments as it is placed on the back of the tablet and mainly captures the background noise of the environment. The CHiME-3 recordings are transcribed using two baseline ASR systems provided by the organizers: one employs Gaussian mixtures to compute acoustic probabilities, while the other uses deep neural networks. Therefore, in the first task (*IWSLT* henceforth) we combine up to 8 different hypotheses, while in the second task (*CHiME-3*) we combine up to 10 different hypotheses. Our goal is to check if, and to what extent, an informed (oracle) ranking at system/segment level can positively contribute to ROVER results.

Ranking	IWSLT (best WER=13.5%)				CHiME-3 (best WER=32.6%)			
	L1	L3	L5	L8	L1	L3	L5	L10
SysO	13.5	12.2	**11.8**	12.1	32.6	30.7	**27.9**	29.4
SegO	**8.9**	10.5	11.4	11.7	**19.5**	21.4	22.4	28.4
InSysO	27.2	19.8	15.1	13.3	40.5	37.5	34.0	29.7
InSegO	33.8	22.9	17.4	13.0	56.3	51.0	42.6	30.9

Table 1: WER[%] ($\downarrow$) results of different hypothesis ranking strategies on IWSLT and CHiME-3 test data.

Table 1 gives the oracle scores at system (SysO) and segment level (SegO) when the best k hypotheses (Lk) are used for the combination (for IWSLT $k = 1, 3, 5, 8$ while, for CHiME-3, $k = 1, 3, 5, 10$). At both levels of granularity, oracle rankings are obtained by measuring the true WER of the candidate transcriptions. As it can be seen in the table, hypothesis combination allows for considerable WER reductions compared to the best individual system, both at system and at segment-level. In particular, on IWSLT, system-level combination results in a WER reduction from 13.5% (best individual system) to 11.8% when combining the best 5 systems. Similarly, on CHiME-3, WER decreases from 32.6% (best individual channel[5]) to 27.9% when the best 5 hypotheses are taken. Note that, in both

---

[5]Henceforth, for CHiME-3 data, the terms "channel" and "microphone" will be interchangeably used.

tasks, system-level combination achieves the best results with only 5 different hypotheses, which is less than all the available ones.[6] This can be explained by observing that all systems/channels contribute to the final voting with the same weight. Therefore, the insertion of the worst hypotheses in the ranked list contributes to worsen rather than to improve global performance.

In both tasks, segment-level combination gives significantly better results than system-level combination. In addition, the WER increases by augmenting the number of hypotheses considered. Actually, being able to correctly select the best hypothesis for each segment (SegO - *L1*) we would obtain the highest performance on both tasks, respectively 8.9% on IWSLT and 19.5% on CHiME-3.

The last two rows of Table 1 give the performance of ROVER when the transcriptions are combined in inverse order, from the worst to the best one, at both system (InSysO) and segment level (InSegO). The poor results achieved, especially at lower combination levels, demonstrate that correctly ranking the input hypotheses is essential to improve ROVER combination results. To address this problem, we investigate two different strategies: ranking by regression (RR) and machine-learned ranking (MLR).

5.2. Ranking by regression (RR)

This strategy exploits a supervised regression model to predict the WER of each transcription at segment-level. Then, it ranks them from the best (lowest predicted WER) to the worst (highest predicted WER). Referring to the structure of Figure 2, for training the QE regressor on a training set with K segments, the training set can be defined as:
$I = \{(S_i^k, H_{i,j}^k, TW_{i,j}^k); 1 \leq k \leq K, 1 \leq i \leq M, 1 \leq j \leq N\}.$
where $TW_{i,j}^k$ stands for the true WER and it is computed by aligning each input hypothesis $H_{i,j}^k$ to the manual reference. The regressor is trained and tuned to predict the true WER following a k-fold cross validation scheme. In the experiments reported in the next sections we will refer to this method as RR1.

For the IWSLT task, where there is only one microphone but several ASR systems, we use a second strategy, which we will call RR2. This method is similar to RR1 but, instead of using all the transcriptions of each segment to train the regressors, it uses only one hypothesis, which is randomly selected. The reason is that in IWSLT the speech signal of a segment is shared among different transcriptions, which makes the signal-based features identical, hence uninformative or even misleading for training the regressors. In this case, the training set can be defined as:
$I = \{(S_i^k, H_{i,j}^k, TW_{i,j}^k); 1 \leq k \leq K, 1 \leq i \leq M, j \in rand(1..N)\}.$
where $rand(1..N)$ outputs a random number between 1 and N. Note that we will not use this second RR strategy for the CHiME-3 task, in which the multiple microphones record different signals for each utterance. In this case, all the signal-based features are different and informative for all the resulting hypotheses.

---

[6]The fact that the best level of combination is the same on both datasets is purely incidental. As we will see in §8.2, the definition of the optimum level of combination for a given dataset is an interesting research direction, which is partially explored also in this work.

5.3. Machine-learned ranking (MLR)

This strategy exploits machine-learned ranking algorithms [17, 49, 21], which are widely used in information retrieval and question answering tasks. Unlike ranking by regression, in which we first predict the WER and then we use it to estimate the rank of each transcription, here we directly predict the ranking. In this approach the labels of the training instances are the true ranks:

$I = \{(S_i^k, H_{i,j}^k, TR_{i,j}^k); 1 \leq k \leq K, 1 \leq i \leq M, 1 \leq j \leq N\}.$

where $TR_{i,j}^k$ is the corresponding "true rank" derived from the true WER value $(TW_{i,j}^k)$. For instance, given two hypotheses ($H_{x,y}^k$ and $H_{z,w}^k$) and their true WERs ($TW_{x,y}^k$ and $TW_{z,w}^k$), we consider $TR_{x,y}^k \preceq TR_{z,w}^k$, if $TW_{x,y}^k \leq TW_{z,w}^k$. Similarly to the RR approach, the ranking models are trained on feature vectors extracted from the signal and the corresponding transcription hypotheses.

MLR, differently from the regression-based strategy, performs a pairwise comparison between the candidates [17]. For any pair of segment transcriptions generated by two or more systems , it processes the corresponding feature vectors and decides to place one transcription ahead of the other. The objective function is designed to minimize the average number of inversions in ranking.

The learning capability of MLR is affected by the level of separability of the instances to be ranked. Sets of hypotheses characterized by large quality differences will favour this learning algorithm, while higher similarity in the hypotheses (and, consequently, a larger presence of ties) will reduce the possibility to return correct ranks. Section 8 provides an initial investigation on the sensitivity of MLR to the number of ties in the training data.

5.4. Features

To train the QE models, from signals and the corresponding automatic transcriptions we extract four sets of features: signal, textual, hybrid and word-based. **Signal** features look at each signal as a whole, trying to capture the difficulty to transcribe it. To this end, they are extracted by analyzing the audio waveform with a window of 20ms length at a frame rate of 10ms. **Textual** features try to model the plausibility (*i.e.* the fluency) of a transcription hypothesis by looking at its content and language model probabilities. **Hybrid** features are intended to characterize the difficulty of transcribing the signal in a more fine-grained way. Also this group is indeed based on signal information, but it also considers as constraint the word time boundaries provided by the decoder at transcription time. These three groups of basic features, correspond to those proposed in [53]. **Word-based** features are new in ASR QE and they are inspired by previous approaches to ASR word error detection [19, 56, 32, 66]. They aim to capture word pronunciation and recognition difficulty, which can be respectively obtained by counting the number of homophones/heteronyms (similar pronunciations) and by computing word-level language model (LM) probabilities. For the former we use a lexicon and for the latter we use n-gram and recurrent neural network language models [51]. These language models are trained on the official data provided by IWSLT and CHiME-3 organizers. Having segments of different length, word-based features are computed by averaging the values of their individual words. The complete list of features is reported in Table 2.

Signal (17)	mean values of 12 Mel Frequency Cepstral Coefficients removing the 0th order coefficient is discarded (12), log energy computed on the whole segments (1), the mean/min/max values of raw energy (3), total segment duration (1).
Textual (10)	number of words (1), LM log probability (1), LM log probability of part of speech (POS) (1), log perplexity (1), LM log perplexity of POS (1), percentage (%) of numbers (1), % of tokens which do not contain only "[a-z]" (1), % of content words (1), % of nouns (1), % of verbs (1).
Hybrid (26)	signal-to-noise ratio (SNR) (1), mean/min/max noise energy (3), mean/min/max word energy (3), (max word - min noise) energy (1), number of silences (#sil) (1), #sil per second (1), number of words (#wrd) per second (1), $\frac{\#sil}{\#wrd}$ (1), total duration of words (D_{wrd}) (1), total duration of silences (D_{sil}) (1), mean duration of words (1), mean duration of silences (1), $\frac{D_{sil}}{D_{wrd}}$ (1), $D_{wrd} - D_{sil}$ (1), standard deviation (std) of word duration (1), std of silence duration (1), mean/std/min/max of pitch[7] (4), number of hesitations (1), frequency of hesitations (1).
Word (22)	POS-tag/score of the previous/current/next words (6), RNNLM probabilities given by models trained on in-domain/out-of-domain data (2), in-domain/out-of-domain 4-gram LM probability (2), number of phoneme classes including fricatives, liquids, nasals, stops and vowels (5), number of homophones (1), number of lexical neighbors (heteronyms) (1) binary features answering the three questions: "is the current word a stop word?"/"is the current word before/after repetition?"/"is the current word before/after silence?" (5).

Table 2: Complete list of the 75 features used to train our ASR QE models.

6. Experimental setting

In this section we describe the data used for our experiments, as well as the metrics and the baselines adopted to evaluate our approach.

6.1. IWSLT

For this task we use the submissions to the IWSLT2012 and IWSLT2013 ASR evaluation campaigns respectively as training and test sets. Both sets are collected from English TED talks dealing with different topics. Six groups participated in IWSLT2012, in which the best performance (12.4% WER) was obtained by NICT[8] group. Two more groups participated in IWSLT2013, which was won again by NICT (13.5%). Tables 3 and 4 provide some basic statistics about this data and participants' WER results. Complete details about each ASR system can be found in [26] and [18].

By selecting the data from two different years, we guarantee that there is no speaker nor topic overlap between training and test sets. In addition, although most of the submissions of 2013 come from the same laboratories of 2012 (except for two teams that did not participate in 2012) the ASR systems are quite different due to the changes and the improvements made

---

13

Attributes	IWSLT2012 (training)	IWSLT2013 (test)
duration (hr)	1h45m	4h50m
# sent	1,124	2,246
# token	19.2k	41.6k
dict. size	2.8k	5.6k
# speakers	11	28

Table 3: Statistics of IWSLT task.

System	IWSLT2012	IWSLT2013
FBK	16.8	23.2
KIT	12.7	14.4
MITLL	13.3	15.9
NAIST	–	16.2
NICT	12.4	13.5
PRKE	–	27.2
RWTH	13.6	16.0
UEDIN	14.4	22.1
Avg.	13.86	18.56

Table 4: IWSLT WER[%] ($\downarrow$) results.

by participants during the course of one year of research. Finally, in this audio dataset, since the recordings were carried out by head-mounted microphones, there is neither reverberation nor background noise apart from the applause breaks, especially in the final part of the talks. However, it happens frequently that the speakers, sometimes non-native ones, introduce spontaneous speech phenomena such as hesitations, repetitions and false starts during their talks. In short, the experiments on this task are conducted as follows:

1. Segmentation of IWSLT2012 (training) and IWSLT2013 (test) data;
2. Feature extraction from the $< signal, transcription >$ training pairs;
3. Training of RR/MLR models;
4. Feature extraction from the $< signal, transcription >$ test pairs;
5. Estimation of the WERs/ranks of the transcribed test segments;
6. Execution of ROVER and concatenation of the resulting outputs;
7. Computation of the WER of the concatenated outputs.

While for the IWSLT2012 audio data manual utterance segmentation is available, for IWSLT2013 the segmentation has to be carried out automatically by each individual ASR system before decoding the audio tracks. Since each system produces a different number of segments, it is necessary to align each automatic segmentation with a reference one in order to share the same utterance time boundaries among the different hypotheses. Although in principle the segmentation given by one randomly chosen ASR system could be taken as reference, we decided to align each segmentation with the reference segmentation provided within the framework of the IWSLT2013 campaign.

6.2. CHiME-3

For the multiple microphones task, we use the publicly available data collected for the 3rd CHiME challenge. In this scenario, six microphones placed on a tablet PC were used to record sentences of the Wall Street Journal corpus, uttered by different speakers in four noisy environments (bus, cafe, pedestrian area, and street junction). Two types of noisy data are provided for this task: *real* data, *i.e.*, speech recorded in four real noisy environments uttered by actual talkers; *simulated* data, *i.e*, noisy utterances generated by artificially mixing clean

14

Attributes	DT05 (training)	ET05 (test)
duration (hr)	2h74m	2h33m
# sentences	1,640	1,320
# words	27.1k	21.4k
dict. size	1.6k	1.3k
# speakers	4	4

Table 5: Statistics of CHiME-3 task.

Channels	DT05	ET05
1-dnn	20.5	32.9
1-gmm	23.4	37.3
3-dnn	20.0	38.8
3-gmm	24.2	40.5
4-dnn	18.8	38.0
4-gmm	21.5	36.5
5-dnn	16.7	32.6
5-gmm	18.7	33.2
6-dnn	16.5	34.4
6-gmm	19.2	34.8
Avg.	19.95	35.9

Table 6: WER[%] ($\downarrow$) of CHiME-3 channels.

speech data with noisy backgrounds. In our experiments we use the `real` subsets. As training data we use *dt05_real* (DT05 henceforth), which is formed by 1,640 sentences uttered by four speakers. As test data we use *et05_real* (ET05), which is formed by 1,320 utterances uttered by four other speakers. In contrast to IWSLT, since each utterance recording shares the same time segmentation across all microphones, here no automatic segmentation is needed.

The organizers of the challenge also provided two baseline ASR systems employing the Kaldi toolkit [57]. One uses the traditional Gaussian mixture model ($* - gmm$) and the other uses a state-of-the-art deep neural network ($* - dnn$). The former is trained with the Kaldi recipe prepared for the previous CHiME challenges [6, 71] and the latter is trained with Karel's setup [70] included in the toolkit.

Tables 5 and 6 respectively show some statistics about the CHiME-3 data and the WER obtained by the baseline ASR systems over each of the 5 different channels (microphones). As mentioned in §5, one of the microphones (index 2) is not used in the evaluation as it is located on the back of tablet PC, mainly to capture the background noise.

Experiments are conducted similarly to the IWSLT task, with the difference that in this case each hypothesis is generated by the two baseline systems (*gmm* and *dnn*) processing each individual microphone audio track ($1, 3, 4, 5, 6$).

In distant speech recognition, it is common to combine the signals recorded by a microphone array in order to generate an enhanced signal. As mentioned in §2, MVDR and DS are two popular enhancement approaches. MVDR processing is included in the baseline system provided by CHiME-3 organizers [4]. This adaptive algorithm minimizes the variance of the resulting signal and it is implemented according to [50]. Similarly, DS [14] applies a fixed weight for each microphone, as described in [38]. For both approaches speaker position is derived from a frame-level computation of the Steered Response Power Phase Transform (SRP-PHAT) that estimates the time difference of arrival between microphone pairs [40, 13], using a 3-dimensional grid. The peaks of SRP-PHAT are then tracked over time using the Viterbi algorithm (see also [46, 7] for implementation details and a comparison of source localization techniques.) We apply both methods to combine the signals from the 5 micro-

phones and then we transcribe the resulting signals by using the two aforementioned baseline ASR systems. This produces four additional channels whose WERs are reported in Table 7.

WER of Enhanced Channels		
Enhancment-ASR	DT05	ET05
mvdr-gmm	20.3	37.1
mvdr-dnn	17.6	33.1
ds-gmm	12.2	23.1
ds-dnn	10.4	20.5

Table 7: WER[%] ($\downarrow$) scores of the enhanced channels.

As it can be seen, DS processing significantly outperforms MVDR on both training and test sets. Nevertheless, we keep also the MVDR hypotheses for combination, since we believe that they provide complementary solutions whose diversity might improve final results.

Comparing the WER scores reported in Tables 4 and 6, it is worth noting that the quality of the CHiME-3 transcriptions is globally lower and, on average, training and test data are more distant. As we will discuss in §7, this can explain why: *i)* the performance achieved by standard random ROVER is lower on CHiME-3 data than on IWSLT, *ii)* adding high quality transcriptions to the combination (*e.g.* those obtained from signal enhancement) results in much larger WER reductions on CHiME-3, and *iii)* ranking methods seem to suffer from the presence of a large number of transcriptions with very similar WER scores.

6.3. Terms of comparison

In both sets of experiments we compare our approach with the standard random ROVER and the two oracles described in §5. In random ROVER the entire transcription candidates obtained from different systems/channels are taken in random order. In the system-level oracle (SysO) we assume to know their true overall rank, which is kept unchanged for all the segments. Moving to a finer granularity, in the segment-level oracle (SegO) we assume to know the exact rank of the ASR hypotheses at the segment level and ROVER is applied segment by segment using the known ranks. The result obtained in this way is our strongest term of comparison.

We consider the last two methods as "oracles" as they rely on external information about the actual system ranks, which is not accessible to the QE-informed approach that we aim to evaluate. Although our first goal is to significantly outperform random ROVER, reducing the performance gap with respect to the two oracles would hence represent an even more convincing measure of success.

6.4. Evaluation metrics

All our results are computed by measuring word error rate. WER is computed by dividing the total number of errors (substitutions, insertions and deletions) by the total number of reference words.

$$WER = \frac{\#substitutions + \#insertions + \#deletions}{\#reference\_words}$$

The regressor parameters are tuned to minimize mean absolute error (MAE) between the predicted WER ($predWER$) and true WER ($trueWER$). For K segments, recorded by M microphones and transcribed by N ASR systems, MAE is computed as:

$$MAE = \frac{1}{K \times M \times N} \times \sum_{k=1}^{K} \sum_{i=1}^{M} \sum_{j=1}^{N} \left| predWER_{i,j}^{k} - trueWER_{i,j}^{k} \right|$$

To tune the parameters of the ranking models we maximize the mean average precision ($MAP@L$) of the predicted ranks at each level of combination ($L = M \times N$). This quantity is defined as follows:

$$MAP@L = \frac{1}{K} \sum_{k=1}^{K} AP^k@L$$

where K is the total number of segments and $AP^k@L$ is the average rank precision of the kth segment. $AP^k@L$ is computed as follows:

$$AP^k@L = \frac{\sum_{l=1}^{L} \frac{P(l)}{l}}{L}$$

where $P(l)$ is the number of correct ranks at level l.

To perform significance tests, we run the matched-pairs test [31] at a significance level of 95% (*i.e.* $p = 0.05$). This method randomly selects subsets of segments processed with two given combination approaches and calculates the corresponding WERs in each subset. If the percentage of agreements between the two approaches is above 95%, then they are not considered as significantly different. In order to simplify the comparison between different combination methods in the result tables, we use the three following symbols:

1. "†" indicates that the WER score obtained after running ROVER with a given ranking method is not significantly different from random ROVER;
2. "•" indicates that the WER score obtained after running ROVER with a given ranking method is not significantly different from the system-level oracle;
3. "⋆" indicates that the WER score obtained after running ROVER with a given ranking method is not significantly different from the segment-level oracle.

Based on the above definitions "†" indicates a negative result, as it is close to the baseline random ROVER, while "•" and "⋆" indicate a positive result, close to the oracles.

6.5. Ranking setup

In order to implement the first ranking strategy, *i.e.* ranking by regression (RR), we rely on the findings of [53], which indicates that extremely randomized trees (XRT) [30] is a suitable regression algorithm for this task. XRT is a combination of several small trees, each trained on the whole data but with a random subset of features. Afterwards, similarly to other ensemble classification methods, an ensemble approach is used for the fusion of the output of all these small trees. For our experiments we use the implementation provided in the Scikit-learn package [55]. The parameters like the number of extra trees, the number of features per tree and the number of instances in the leaves are tuned to minimize MAE, using k-fold cross-validation on the training set.

For the second strategy, *i.e.* machine learned ranking (MLR), we adopt the solution proposed in [39], which identifies in the random forest algorithm [15, 21] the best ranking model for this kind of tasks. In particular, we use the implementation provided as an option in the RankLib library.[9] The main parameters, such as the number of bags, the number of trees per bag and the number of leaves per tree are tuned to maximize MAP, using k-fold cross-validation.

All models are trained with three different feature settings: *i)* signal, textual and hybrid features, indicated in the tables below with "*+B*"; *ii)* word-based features, indicated with "*+W*"; and *iii)* the combination of both, indicated with "*+BW*".

7. Results

For both IWSLT and CHiME-3 tasks, we run our QE-informed ROVER approach at different levels of combination, that is with different numbers of ASR components contributing to the fusion process. This number is always higher than 3, since a lower value would be meaningless when majority voting is applied. Therefore, the combination level for IWSLT ranges in the interval $[L3, L8]$, while for CHiME-3 it ranges in $[L3, L14]$.

It's worth remarking that the results reported in this work are not comparable with the official ones of CHiME-3, since here we assume that neither the ASR systems nor their confidence scores are accessible. Instead, the official submissions (like our own ones described in [38]) take advantage of a combination of acoustic model adaptation, confidence-based data selection and multi-pass recognition as well as signal processing and language model rescoring techniques. Here, we do not consider these techniques and we apply our QE approach to the first pass decoding output of the baseline systems. This is to confirm that our method is general enough to be applied in a variety of "black-box" conditions.

7.1. IWSLT

Preliminary analysis. To test the performance of our method in a controlled setting, we first run QE-informed ROVER on the training set (IWSLT2012) using 4-fold cross validation. Data partitioning is done with regard to the speakers in order to guarantee that there is no

---

[9]http://sourceforge.net/p/lemur/wiki/RankLib

Ranking	L3	L4	L5	L6	Avg. Impr.
Random	13.4	11.8	12.3	11.8	0.0
SysO	11.4	9.3	9.6	9.5	-2.4
SegO	8.0	7.9	8.2	9.1	-4.0
RR1+B	11.5	10.2	9.9	9.8	-2.0
RR1+W	13.2	10.6	10.1	9.8	-1.4
RR1+BW	12.1	10.3	10.0	9.8	-1.8
RR2+B	11.3	10.3	9.9	9.8	-2.0
RR2+W	12.1	10.3	10.0	9.8	-1.8
RR2+BW	11.2	9.9	9.8	**9.6**	-2.2
MLR+B	10.7	9.8	9.7	**9.6**	-2.4
MLR+W	10.7	**9.7**	9.7	**9.6**	-2.4
MLR+BW	**10.6**	9.8	**9.6**	**9.6**	-2.4

Table 8: WER[%] ($\downarrow$) of different ranking methods on IWSLT2012 using 4-fold CV.

speaker overlap between training and test folds. Moreover, the partitioning is done in such a way that each instance occurs only once in each test part. Thus, by collecting the results from all folds, we obtain the performance on whole training data reported in Table 8.

The first three rows of the table show the results of ROVER when the input hypotheses are randomly ordered at system level, and those achieved by the two oracles previously described (SysO and SegO). Even if random ROVER achieves better results than the individual IWSLT2012 participants (see Table 4), it is quite far from the oracles. SysO, for which the overall ranking of the ASR systems is known and is kept unchanged for all the segments, is significantly better, with an average WER reduction of 2.4% over random ROVER. As expected, the best performance is achieved by the segment-level oracle with an average WER reduction of 4.0%. These differences confirm that running ROVER without an optimal order of the hypotheses may limit the gain achievable by system combination. Apart from that, both oracles reach the lowest WER score at combination level $L4$. This indicates that the majority of aligned words in the confusion sets of the word network built by ROVER falls in the top four hypotheses. Inserting more alternative words from less accurate hypotheses at $L5$ and $L6$ has a negative impact on majority voting.

Ranking by regression, using all the training transcriptions and the basic features (RR1+B) already outperforms random ROVER significantly but remains far from the oracles. Note that basic features include signal, textual and hybrid features from Table 2. Slightly worse performance is obtained when using the word-based features (RR1+W) and the union of the features (RR1+BW), suggesting that the word features are not particularly useful in this setting.

Although the various ASR systems provide different hypotheses for each decoded segment, the corresponding signal feature vectors are the same and, therefore, the trained regressor could be misled by the co-occurrence of similar feature values with different training labels. To overcome this problem, as explained in §5.2 we trained regression models

(called RR2 in Table 8) by using only one random transcription for each segment. This approach, using both basic and word-based features (RR2+BW), yields the best performance among the RR methods (on average, the WER is reduced by 2.2%). Interestingly, both RR1 and RR2 reach the best performance at combination level $L6$. This behavior, differently from the oracle conditions, indicates that the majority of the words in each confusion set of the ROVER word network is not concentrated on the top "automatically" ranked positions but is more uniformly distributed over all combination levels. Machine-learned ranking with the basic features (MLR+B), word-based features (MLR+W) and their combination (MLR+BW) always exhibits the largest WER reductions over random ROVER (-2.4% on average). Its results are not only consistently better than RR, but also close to system-level oracle (SysO).

In summary, this preliminary analysis suggests that: *i)* regardless of the QE method and the features used, QE-informed ROVER outperforms random ROVER with a large margin, *ii)* among all the QE strategies, MLR+BW shows the best performance at most of the levels, and *iii)* though still far from the strongest term of comparison (SegO), its results are competitive with those of the system-level oracle. As shown in the last column of Table 8, the average improvement of the two methods are almost the same (-1.3% vs. -1.1%).

Test. Using the QE models trained on IWSLT2012, we carry out the same set of experiments on the test data. The corresponding results are reported in Table 9. Since in IWSLT2013 there are 8 teams who submitted their transcriptions, in this case the combination level varies in $[L3, L8]$ interval.

Ranking	L3	L4	L5	L6	L7	L8	Avg. Impr
Random	14.6	13.7	13.2	12.8	12.7	12.4	0.0
SysO	12.2	11.7	11.8	11.9	12.1	12.1	-1.3
SegO	10.5	11.0	11.4	11.6	11.7	11.7	-1.9
RR1+B	13.9	13.1	12.6	12.4	12.4	12.3 † ●	-0.4
RR1+W	14.0	13.0	12.5	12.2	12.3 ●	12.3 † ●	-0.5
RR1+BW	14.0	13.0	12.5	12.2	12.3 ●	12.3 † ●	-0.5
RR2+B	13.8	13.0	12.6	12.4	12.3 ●	12.3 † ●	-0.5
RR2+W	14.2	13.1	12.7	12.4	12.5 †	12.4 † ●	-0.3
RR2+BW	13.7	12.8	12.4	12.2	**12.2 ●**	**12.2 † ●**	-0.6
MLR+B	12.9	12.4	12.3	12.1 ●	12.3	**12.2 † ●**	-0.9
MLR+W	**12.4 ●**	**12.1**	**12.0**	12.0 ●	**12.2 ●**	**12.2 † ●**	-1.1
MLR+BW	**12.4 ●**	**12.1**	**12.0 ●**	**11.9 ● ⋆**	**12.2 ●**	**12.2 † ●**	-1.1

Table 9: WER[%] ($\downarrow$) of different ranking methods on IWSLT2013. "†" = the result is not statistically different from random ROVER; "●" = the result is not statistically different from SysO; "⋆" the result is not statistically different from SegO.

Although the overall trend is similar to the one observed on the training data, the improvement margins are smaller. The reason for this unsurprising difference is the mismatch between training and test data. In fact, in our preliminary experiments, the same set of ASR

systems (6 systems) was used both in training and test. Here, instead, the ASR systems are different in number and nature (6 systems from IWSLT2012 for training, 8 systems from IWSLT2013 for test). Nevertheless, in many columns of Table 9 the WER reduction is large and significant. The best result (MLR+BW at $L6$), in particular, is not only better than random ROVER, but also not statistically different from the strongest oracle (SegO). Also on the test set, MLR seems to work better than RR in ranking the input hypotheses. One possible reason is the higher reliability of pairwise comparisons compared to considering numeric scores that reflect independent quality predictions.

So far, we showed that: *i)* we can improve random ROVER using ASR QE for ranking the inputs, and *ii)* with an appropriate QE approach and an efficient set of features, we can approach the strong system-level and segment-level oracles. The scores reported in Table 9, however, only provide a global picture that might hide interesting details such as larger gains in specific conditions that are particularly favorable for QE-based ranking.

To better investigate this aspect, in Figure 3, we analyse the performance on different groups of segments characterized by different levels of diversity. By diversity, we refer to the level of disagreement between ASR systems. It is computed as the difference between the maximum and minimum WERs among the transcriptions of a given segment. With this definition, we divide the segments into 10 groups (X-axis in Figure 3). The first group consists of the segments whose transcriptions diversity is lower than 10% (*i.e.* the WER difference between the best and the worst transcription is less than 10%). In the second group, the diversity is between [10%,20%] and so on. The figure reports the performance of ROVER using SegO, SysO and MLR+BW over each diversity group, as well as the proportion of segments belonging to each group (black-dotted line). As it can be seen, most of the segments have low diversity (less than 20%), meaning that the ASR systems return similar transcriptions for them. It is interesting to note how the relative differences between the three systems are affected by hypothesis diversity. For low values, in which ranking is intuitively more difficult and less accurate, MLR+BW's performance is less reliable and close to SysO. For high diversity values (*i.e.* above 70%), in which the ranker is able to order the components more precisely, MLR+BW outperforms SysO halving the gap that separates it from SegO in the last two bins. This interesting trend is hidden by the fact that the IWSLT data are characterized by low diversity in the transcriptions (highly-diverse hypotheses can be found for less than 7% of the segments). This, in turn, results in global WER scores where large gains on few segments are smoothed by small gains on many segments. We hypothesize that, though suitable to evaluate our approach, the IWSLT scenario is not the ideal one to fully appreciate its potential. In the next set of experiments we apply QE-informed ROVER to the CHiME-3 data, in which the diversity among the microphone channels is higher. If our hypothesis is correct, then larger WER gains over SysO and further reductions in the distance from SegO will be observed.

7.2. CHiME-3

Preliminary analysis. Differently from the IWSLT task, where the transcriptions come from multiple ASR systems, in CHiME-3 the transcriptions come from multiple channels and two ASR systems. Also in this task we first train the QE models on the training set (DT05)
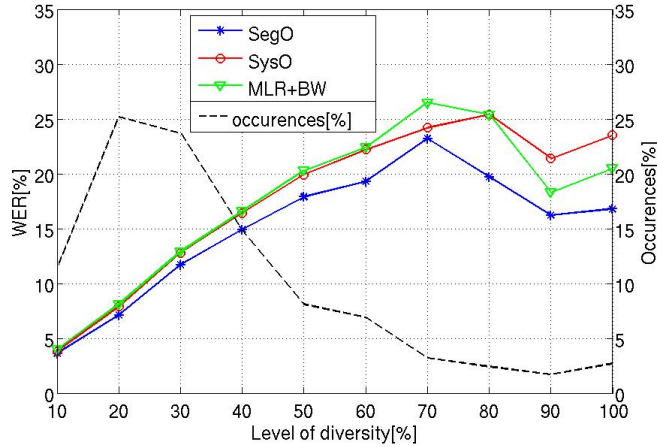
Figure 3: Performance as function of hypothesis diversity ($div = MAX_{WER}[\%] - MIN_{WER}[\%]$) in IWSLT.

using 8-fold cross validation. The partitioning is done to avoid speaker or sentence overlaps between training and test folds. As mentioned in §6.2, for each utterance there are five signals recorded by the microphones, plus two enhanced signals. Each signal is transcribed using both GMM and DNN acoustic models. Hence, 14 hypotheses are computed for each segment and, consequently, the combination levels range in the interval $[L3,L14]$.

Ranking	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	Avg. Impr.
Random	14.3	13.4	12.7	12.4	12.1	11.9	11.9	11.8	11.8	11.8	11.8	12.2	0.0
SysO	11.0	10.9	10.3	10.4	10.6	10.6	10.9	10.7	10.9	10.9	11.2	11.5	-1.5
SegO	6.8	7.1	7.5	7.8	8.3	8.6	9.2	9.4	10.0	10.4	10.9	11.2	-3.4
RR1+B	10.6	10.3	10.0	9.9	10.0	10.1	10.3	10.4	10.8	11.0	12.0	12.0	-1.7
RR1+W	12.3	11.4	10.9	10.7	10.7	10.7	10.8	11.1	11.3	11.7	12.0	12.1	-1.0
RR1+BW	10.3	10.0	9.7	9.7	9.9	**9.9**	**10.2**	10.4	10.7	**10.0**	11.7	**11.9**	-2.0
MLR+B	10.0	9.7	9.6	9.6	9.8	10.0	**10.2**	10.4	**10.5**	10.7	**11.3**	**11.9**	-2.0
MLR+W	10.7	10.5	10.3	10.3	10.3	10.5	10.7	11.0	11.3	11.5	11.6	12.0	-1.4
MLR+BW	**9.8**	**9.5**	**9.5**	**9.5**	**9.7**	**9.9**	**10.2**	**10.3**	**10.5**	10.8	11.5	**11.9**	-2.1

Table 10: WER[%] (↓) of different ranking methods on the CHiME-3 training set (DT05) using 8-fold CV.

In this task we do not take into account the RR2 approach. Indeed, since the microphones that record the signal are positioned in different places over the tablet PC, each of them captures a different signal, thus making all the signal-based features different and informative.

The results reported in Table 10 show that, similar to the IWSLT task, the best performance is achieved in most of the cases by MLR+BW (with an average WER reduction of 2.1% over random ROVER). Unlike IWSLT, most of our results outperform the SysO oracle, which always uses the enhanced channels as the first input. The best WERs are usually

obtained at the lower levels (*i.e.* 9.5% at $L4$, $L5$ and $L6$). This can be explained by the large gap between the best channels ($ds - dnn$ and $ds - gmm$, respectively obtaining 10.4% and 12.2% WER) and the other channels (obtaining from 16.5% to 24.2% WER). By adding more hypotheses, worse transcriptions are considered in the combination and, consequently, majority voting makes more mistakes. At $L13$ and $L14$, indeed, our results remain slightly worse than SysO.

An interesting achievement with this approach is the significant improvement of hypothesis-level combination with respect to the signal-level combination. The top result obtained by MLR+BW (9.5%) is indeed better than the best one reported in Table 7 for signal-level combination (10.4% with $ds - dnn$). In contrast with previous works suggesting that hypothesis-level combination does not yield any significant improvement on top of signal-level combination [65], our results show that better final results can be obtained through an accurate ordering of the inputs.

Test. Table 11 shows the WER results achieved when QE models are trained on the DT05 and then used to predict the quality of the ET05 transcriptions. The observations of our preliminary analysis seem to be confirmed. Although trained and tested on different data, QE-informed ROVER consistently outperforms random ROVER and the SysO oracle at all levels except for $L14$, where the WER variations achieved by our best system are not statistically significant. This supports our hypothesis that, if the diversity among the components is high enough, then the proposed segment-level QE-informed ROVER can lead to better results than the system-level oracle. SegO performance is still significantly better at all levels, especially the lower ones where the transcriptions obtained from the best enhanced channels are hard to improve with hypotheses coming from the other channels.

It is important to remark that the QE-informed ROVER also significantly improves the performance of the enhanced channels when combining less than eight transcriptions. In fact, as shown in Table 7, the enhanced $ds - dnn$ channel achieves 20.5% WER on ET05, while RR1+BW reduces the error down to 19.1% at $L5$. As mentioned before, probably due to the high performance difference between the enhanced and raw channels, we do not observe the same gain at higher levels.

Differently from IWSLT, where there is an increment of around 2 WER points when moving from the preliminary analysis to the test results, in CHiME-3 we observe a larger degradation ($\sim$10.0% WER on average). This is not surprising if we look at the differences in WER between training and test data in the two tasks (Tables 4 and 6), which indicate that CHiME-3 data is more heterogeneous. However, though working in a more complex scenario, the QE-informed ROVER is still able to achieve competitive results. In terms of learning approach, in most of the cases the best performance on CHiME-3 is obtained by RR1. This is in contrast with the IWSLT results, where MLR always performed better than RR. The main reason for this is related to the fact that in CHiME-3, apart from the enhanced channels, the other channels are quite similar in performance and, overall, they generate transcriptions of low quality with similar or equal WERs (see Table 6). For MLR, this results in a large number of ties when computing the pairwise comparisons. The impact on MLR performance with a proper management of ties is addressed in §8.1.

Ranking	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	Avg. Impr.
Random	28.7	27.4	27.2	26.8	26.5	26.4	26.2	26.2	26.1	26.1	26.2	26.3	0.0
SysO	22.7	22.0	21.6	21.1	21.9	22.0	24.0	24.1	24.4	24.9	25.5	25.8	-3.3
SegO	14.8	15.4	16.2	17.0	18.1	18.7	19.7	20.3	21.1	21.9	22.8	23.6	-7.4
RR1+B	20.4	**19.5**	19.5	20.1	20.2	20.6	21.1	21.7	22.2	**22.9**	24.2	25.8†•	-5.1
RR1+W	22.2•	21.1	20.3	20.3	20.5	21.0	21.3	22.0	22.9	24.0	24.9	26.0†•	-4.5
RR1+BW	20.0	**19.5**	**19.1**	**19.5**	**19.7**	**20.3**	**20.7**	**21.4**	**22.1**	**22.9**	**23.9**	25.8†•	-5.4
MLR+B	20.4	19.9	19.6	20.0	20.3	20.6	21.2	21.5	22.5	23.3	24.9	25.8†•	-5.0
MLR+W	22.4•	21.4	20.8	20.7	21.1	21.5	22.0	22.6	23.2	24.0	25.2	25.9†•	-4.1
MLR+BW	**19.8**	**19.5**	19.5	19.7	20.2	20.4	20.9	21.5	22.2	23.4	24.9	**25.7**•	-5.2

Table 11: WER[%] ($\downarrow$) of different ranking methods on the CHiME-3 test set (ET05). "†" = the result is not statistically different from random ROVER; "•" = the result is not statistically different from SysO; "⋆" the result is not statistically different from SegO.
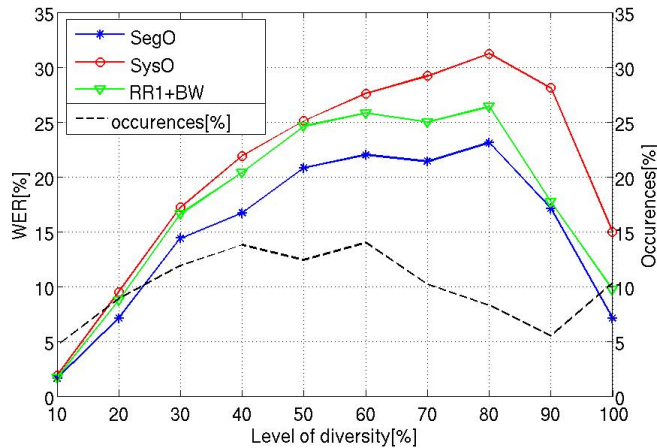

Figure 4: Performance as function of hypothesis diversity in CHiME-3.

Figure 4 shows the performance analysis with regard to the diversity groups in the CHiME-3 task. Comparing the curves in Figures 3 and 4, we observe that the distribution of the CHiME-3 transcriptions with respect to the different levels of diversity (black-dotted lines) is more uniform than in IWSLT. This is mostly due to the presence of the enhanced channels that perform significantly better than the raw ones, hence they enlarge the gap between the best and the worst transcriptions. In terms of performance, this uniform distribution allows the QE-informed ROVER to outperform SysO also for small values of diversity ($\sim$30%) in CHiME-3. Increasing the level of diversity, our method significantly gains in performance over SysO and also approaches SegO at the diversity value of 90%.

In this section we have shown that: *i)* QE-informed ROVER is able to achieve better performance than standard ROVER on two very different datasets, *ii)* our approach also outperforms the system-level oracle in the CHiME-3 task, and it obtains competitive results

with the enhanced channels, *iii)* the extent of our gains depends on the level of diversity and accuracy of the transcriptions and *iv)* the smallest WER scores are obtained by combining a limited number of transcriptions. One can question how the optimal number of transcriptions can be found in a real unsupervised scenario where the references to compute the WER are not available. We will answer this question in the next section, after analyzing the impact on performance of tied ranks.

8. Practical issues

Despite the positive results discussed so far, our analysis still leaves two important questions open:

1. Why machine-learned ranking (MLR+BW) shows better performance on IWSLT, whereas in CHiME-3 ranking by regression (RR1+BW) generally works better?
2. Considering that the best performance is always obtained at a task-specific level of combination, how can we predict this level?

In this section we address these two practical issues, respectively by: *i)* analyzing the impact of tied ranks (transcriptions with identical WERs) in the CHiME-3 training data and *ii)* exploring a classification method to identify the optimum combination level.

8.1. Tied ranks

In the previous section we observed that the highest results on IWSLT and CHiME-3 data are respectively obtained by MLR and RR. This raises the following practical issue: how to find a unique, best performing strategy, suitable for the general case?

One reason for the observed difference can be found in the way the QE models are trained. In RR they are trained to separately predict the WER of each transcription, while in MLR they are trained to predict relative ranks through pairwise comparisons. Training sets characterized by a large number of ties (transcriptions of the same segment with identical WERs but different quality) may influence the performance of MLR more than RR.[10]

Table 12 reports the average percentage of ties for each dataset, showing that IWSLT has less ties than CHiME-3 both in the training and test sets. This seems to contradict the plots of Figure 4, which indicate a higher diversity in the CHiME-3 data (compared to IWSLT, instances are more evenly distributed across all diversity levels). Such diversity, however, is only due to the presence of the enhanced channels. These are far better than the other ones, whose performance is uniformly lower and leads to a large number of similar or identical transcriptions. Moreover, the CHiME-3 training set includes 8% more ties than the test set. This suggests that the presence of a large number of ties in the training set of CHiME-3 can be critical and, in turn, it can determine the lower results achieved by MLR.

---

[10]This problem is widely explored in the information retrieval field, where ties are either arbitrarily broken [29] or managed with *ad-hoc* strategies [79].

	IWSLT	CHiME-3
Train	38.4%	58.9%
Test	40.5%	50.6%

Table 12: Percentage of ties in each dataset.

To validate this hypothesis, in the following experiments we break the ties in the training set by looking at the global performance of each system/channel on the same data.[11] In particular, if two systems, A and B, achieve the same WER on a given training segment, and system A has shown to perform better than system B on the whole training set, then the hypothesis suggested by A will be prioritized when breaking the ties.

Table 13 shows how performance changes if the ties in the training data are broken using such prior knowledge. In IWSLT, the MLR approach trained on untied ranks ($MLR + BW + Untied$) achieves minor improvements. This is due to the fact that: *i)* the number of ties is not so high to represent a critical issue as we saw in Table 12, and *ii)* the small WER differences between systems and combination levels reduce the room for improvement. In CHiME-3, instead, the model learned from the untied training set ($MLR + BW + Untied$) significantly outperforms $MLR + BW$ at all levels of combination and also outperforms $RR1 + BW$ at most of the levels. The considerable WER result of 18.6% obtained by this method at $L4$, which also corresponds to a relative improvement of 9.6% over the best enhanced channel, represents our best result on CHiME-3. This confirms the validity of our intuition about the negative impact of tied ranks on MLR performance. The largest improvement is obtained when combining all the transcriptions ($L14$). This is not surprising because this is the condition where more ties occur. Interestingly, at this level, the results of QE-informed ROVER improve up to the point that they are no longer statistically different from the strong segment level oracle (see Table 11).

Similarly to what we observed in the previous experiments, the best results are obtained at low levels of combination. This calls for solutions to automatically find (or at least approximate) the optimum level, which is the problem investigated in the next section.

8.2. Optimum level of combination

Table 13 shows that, after breaking the ties, the optimum level of combination in IWSLT (with 8 components) is either $L4$ or $L6$, while in CHiME-3 (with 14 components) it is $L4$. A *post-hoc* comparison of the results achieved by each level on the test set, however, is only made possible by the availability of references to compute final WER scores. This observation raises a new practical issue: how can we predict the optimum level of combination in a real condition in which we have no access to the reference transcripts?

The problem of finding a dataset-specific stopping criterion, as an alternative to the simple and risky "take-all" strategy, is well motivated. Although it seems less important

---

[11]This information is similar to the prior knowledge that the SysO oracle acquires on the training data and exploits to rank the test hypotheses. Here, however, we fairly operate only on the training set.

IWSLT	L3	L4	L5	L6	L7	L8	-	-	-	-	-	Avg. Impr.
RR2+BW	13.7	12.8	12.4	12.2	12.2•	**12.2**†•	-	-	-	-	-	-0.6
MLR+BW	12.4•	12.1	**12.0**•	**11.9**•★	12.2•	**12.2**†•	-	-	-	-	-	-1.1
MLR+BW +Untied	**12.3**	**11.9**•	**12.0**•	**11.9**•★	12.1•	**12.2**†•	-	-	-	-	-	-1.2

CHiME-3	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	Avg. Impr.
RR1+BW	20.0	19.5	**19.1**	**19.5**	**19.7**	**20.3**	**20.7**	21.4	22.1	22.9	23.9	25.8†•	-5.4
MLR+BW	19.8	19.5	19.5	19.7	20.2	20.4	20.9	21.5	22.2	23.4	24.9	25.7•	-5.2
MLR+BW +Untied	**19.3**	**18.6**	19.2	**19.5**	20.0	**20.3**	20.8	**21.2**	**21.8**	**22.5**	**23.2**	**24.0**★	-5.8

Table 13: WER[%] ($\downarrow$) results when the ties are broken using prior knowledge. "†" = the result is not statistically different from random ROVER; "•" = the result is not statistically different from SysO; "⋆" the result is not statistically different from SegO.

for IWSLT, where final WER scores are quite similar for all the levels, a method to avoid entering harmful inputs into the ROVER combination can significantly change the results in CHiME-3.

Several sub-optimal solutions can be applied to address this problem. The simplest one is the random choice, which is acceptable in situations where hypothesis quality is homogeneous (IWSLT) but, as we will show below, it can be inadequate when the variability is higher (CHiME-3). Another option is to determine the optimal level in the training set and apply it to the test data. This, however, would not be efficient when the number of components varies between training and test, as in the case of the IWSLT task. A better solution is suggested by the findings of [2], which demonstrates the strong dependency between the performance of ASR system combination methods and the diversity and quality of the components. Following this intuition, we explore a classification-based approach, in which a binary classifier is trained to learn whether a combination level is appropriate (labeled as *1*) or not (labeled as *0*). Based on the predictions of the classifier, for each given segment we select the appropriate level. In case of finding multiple optimum levels, we rely on the confidence score of the classifier.

In the following, we explain this classification task in detail, focusing on the data, features, classifiers and evaluation metric used in our final round of experiments.

Data. To prepare the training data, for each segment we first run the proposed segment-based QE-informed ROVER at all combination levels. Then, label *1* is assigned to the level(s) with the smallest WER. It is important to note that, for some segments, different levels may result in the same smallest WER score. When this happens, multiple levels are labeled as *1* thus resulting in skewed label distributions. This is the case of IWSLT data, especially the test set, in which more levels have identical scores compared to CHiME-3 (the percentage of positive examples in the two test sets is respectively 77.0 and 57.3). Looking at

the results in Table 13 this is not surprising, since the WER difference between the best and the worst level is minimal (0.4% for $MLR + BW + Untied$) compared to CHiME-3 (5.4%). Moreover, looking at the distributions in Figures 3 and 4, we notice that the majority of the IWSLT segments have transcriptions with small diversity. These considerations indicate a higher probability to find transcriptions of similar quality in the IWSLT data and, in turn, a more skewed label distribution when training our classifier.

Features. The features for this classification task are extracted by using the confusion networks generated by ROVER at each combination level. In particular, we compute:

1. Overall diversity of the level, computed by using the method proposed in [2];
2. Levenshtein distance between the first and the last hypothesis;
3. Avg. Levenshtein distance between the first hypothesis and the others;
4. Avg. Levenshtein distance between each hypothesis and the next one;
5. Avg. Levenshtein distance between each hypothesis and the final combination;
6. Mean/min/max predicted WER obtained by RR methods among the hypotheses.

Except for the first feature, the others are quite simple to compute. The approach described in [2] defines the *diversity* of the set of ASR systems to combine as "the average spread of the individual ASR predictions around the average prediction". Given a ROVER WTN (see §3) formed by I confusion slots and built by combining M ASR systems, we define as $\mathbf{w}_i^m$ ($1 \leq m \leq M, 1 \leq i \leq I$) a one-hot binary vector, of dimension equal to the dictionary size, having all 0 components except for the one corresponding to the word furnished by system m in the ith confusion slot. First, the voting scheme of ROVER averages the $\mathbf{w}_i^m$ across the M ASR systems, as follows:

$$\mathbf{w}_i^{avg} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{w}_i^m$$

In this way $\mathbf{w}_i^{avg}$ contains the frequencies of the words in the ith confusion slot. Then, as seen in §3, to provide the combined recognized string, ROVER outputs the word with the highest frequency in each slot i. The diversity measure of M ASR hypotheses being combined is defined with the following Equation [2]. Note that, $\| \ \|_2^2$ represents the vector $L2$ norm.

$$Diversity = \frac{1}{I \times M} \sum_{i=1}^{I} \sum_{m=1}^{M} \frac{1}{2} \| \mathbf{w}_i^{avg} - \mathbf{w}_i^m \|_2^2$$

Classifiers. We experiment with Support Vector Machine (SVM) [47], Random Forest [15] and BayesNet [28] classifiers trained in binary mode. Parameters are tuned with 5-fold cross validation on the training set.

Evaluation metric. Balanced accuracy, which is particularly appropriate in case of skewed distributions, is used as evaluation metric to optimize hyper-parameters and select the best classification algorithm. The cross validation results reported in Table 14 show that BayesNet outperforms the other classifiers and, by a large margin, also the 50.0% balanced accuracy reachable with a baseline system relying on majority voting. One possible explanation for the success of BayesNet could be in its higher capability to work with a limited number of features (8) and a large number of instances (4496 for IWSLT and 13476 for CHiME-3). Comparing the performance obtained on the two datasets, the fact that CHiME-3 results are better than the IWSLT ones suggests that the more skewed distribution of the IWSLT labels penalizes the classifier.

Task	IWSLT	CHiME-3
Classifier	Balanced Acc. IWSLT2012	Balanced Acc. DT05
SVM	52.8	66.1
Random Forest	58.7	71.5
BayesNet	**65.5**	**72.0**

Table 14: Performance of different binary classifiers used to find the optimum level of combination.

In light of its higher balanced accuracy, the BayesNet classifier has been selected to predict the best level of combination for each segment in the test data. On both tasks the results outperform those of the sub-optimal strategy based on random selection of the best combination level. While on IWSLT the WER reduction is unsurprisingly minimal (from 12.1% to 12.0%), on CHiME-3 the gain is much larger (from 20.9% to 18.8%).[12] More interestingly, besides outperforming random selection, our classifier provides predictions that closely approximate the performance of the best levels shown in Table 13 (11.9% for IWSLT and 18.6% for CHiME-3).

Overall, these results indicate that: *i)* the optimal combination does not only depend on the quality of the hypotheses and the reliability of their ranking but also on their diversity (this is in line with the discussion in §7.2 and with the results reported in [2]); *ii)* the relative importance and contribution of these aspects can be learned from data, and *iii)* the resulting models can effectively support our QE-informed ROVER approach in real operating conditions where reference transcripts are not available.

9. Conclusions

We investigated the use of ASR quality estimation as a way to address some potential drawbacks of the current ROVER-based hypothesis combination strategies. These concern the influence of three main factors on the final result: *i)* the availability of *confidence information* about the inner workings of the ASR systems producing the transcriptions, *ii)*

---

[12]The baseline sub-optimal scores are obtained by averaging the results obtained from five iterations of the random selection process.

the *order* in which the input hypotheses are used to initialise the fusion process and *iii)* the *granularity* of the combined transcriptions. Our QE-based approach overcomes the first limitation by informing ROVER with automatic and confidence-independent quality predictions. This allows us to operate in "black-box" application scenarios in which the hypotheses to be combined come from unknown systems. The reliability of our predictions is exploited to overcome the second limitation by sorting the input hypotheses and prioritizing the most accurate ones to support ROVER's voting scheme. The third limitation is tackled by feeding ROVER with reliable rankings at segment-level instead of considering long utterance transcriptions whose quality can considerably vary at local level. This allows our QE-informed ROVER to better exploit the local diversity of the combined hypotheses.

By extending previous experiments to a new scenario, we analyzed the performance of different ranking strategies on two different tasks. Working on the automatic transcription of TED talks from the IWSLT evaluation campaign, we experimented with the combination of outputs from a *single microphone and multiple systems*. Working with the automatic transcription of Wall Street Journal articles from the CHiME-3 campaign, we tested our capability to combine outputs from *multiple microphones and multiple systems*.

In all tasks, our approach significantly outperforms the standard random ROVER, which is fed with entire audio transcriptions presented in random order (absolute WER improvements range from 0.5% to 7.3%). More interestingly, our best results are competitive with those achieved by two strong oracles, which are informed about the true ranks of the input hypotheses respectively at system and segment level.

To make our analysis more complete, we studied performance variations as a function of the diversity of the combined hypotheses and, finally, we addressed two practical issues: *i)* dealing with tied ranks and *ii)* automatically finding the optimum number of hypotheses that have to be combined for each input segment. Our solutions respectively allowed us to isolate a unique, best performing ranking strategy and to deploy it for a competitive and fully automatic method to inform ROVER with QE predictions.

Our results open the door to new applications for ASR QE. Recently, we started to work on the unsupervised adaptation of deep neural networks for acoustic modeling, using segment-level WER prediction to regularize the objective function computed on the adaptation data. ASR QE can also be used to select audio segments with *good* transcriptions to be inserted either in adaptation or in the training sets. In an MDM scenario QE can be employed to select the best microphone (similarly to what we did with CHiME-3 data) or to train improved acoustic models by filtering clean data with different predefined impulse responses. In this context, QE-informed ROVER has been applied to improve the performance given by a more recent CHiME-3 baseline [36, 25].

References

[1] Abida, K., Karray, F., Abida, W., 2011. cROVER: Improving ROVER using Automatic Error Detection, in: Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic. pp. 1753–1756.

[2] Audhkhasi, K., Andreas M., Z., Panayiotis G., G., Shrikanth S., N., 2014. Theoretical Analysis of Diversity in an Ensemble of Automatic Speech Recognition Systems. IEEE/ACM Transactions on Audio, Speech, and Language Processing 22, 711–726.

[3] Bach, N., Huang, F., Al-Onaizan, Y., 2011. Goodness: a Method for Measuring Machine Translation Confidence, in: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), The Association for Linguistics, Portland, Oregon, USA. pp. 211–219.

[4] Barker, J., Marxer, R., Vincent, E., Watanabe, S., 2015. The third 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines, in: Proc. of the 15th IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Scottsdale, Arizona, USA. pp. 1–9.

[5] Barker, J., Vincent, E., Ma, N., Christensen, H., Green, P., 2013a. The PASCAL CHiME Speech Separation and Recognition Challenge. Computer Speech & Language 27, 621–633.

[6] Barker, J., Vincent, E., Ma, N., Christensen, H., Green, P., 2013b. The PASCAL CHiME Speech Separation and Recognition Challenge. Computer Speech & Language 27, 621–633.

[7] Blandin, C., Ozerov, A., Vincent, E., 2012. Multi-source tdoa estimation in reverberant audio using angular spectra and clustering. Signal Processing 92, 1950–1960.

[8] Boersma, P., Weenink, D., 2005. Praat: Doing Phonetics by (Version 4.3.01). Retrieved from http://www.praat.org/ .

[9] Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., Specia, L., 2013. Findings of the 2013 Workshop on Statistical Machine Translation, in: Proc. of the 8th Workshop on Statistical Machine Translation (WMT), Sofia, Bulgaria. pp. 1–44.

[10] Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., Tamchyna, A., 2014. Findings of the 2014 Workshop on Statistical Machine Translation, in: Proc. of the 9th Workshop on Statistical Machine Translation (WMT), Association for Computational Linguistics, Baltimore, Maryland, USA. pp. 12–58.

[11] Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., Turchi, M., 2015. Findings of the 2015 Workshop on Statistical Machine Translation, in: Proc. of the 10th Workshop on Statistical Machine Translation (WMT), Association for Computational Linguistics, Lisbon, Portugal. pp. 1–46.

[12] Bougares, F., Del'eglise, Paul, E., Yannick, Rouvier, M., 2013. LIUM ASR System for Etape French Evaluation Campaign: Experiments on System Combination using Open-source Recognizers, in: Proc. of the 16th International Conference on Text, Speech, and Dialogue (TSD), Pilsen, Czech Republic. pp. 319–326.

[13] Brandstein, M., Ward, D., 2001. Microphone Arrays: Signal Processing Techniques and Applications. Springer.

[14] Brandstein, M.S., Silverman, H.F., 1997. A robust method for speech signal time-delay estimation in reverberant rooms, in: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 375–378 vol.1.

[15] Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32.

[16] Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., Specia, L., 2012. Findings of the 2012 Workshop on Statistical Machine Translation, in: Proc. of the 7th Workshop on Statistical Machine Translation (WMT), Montr'eal, Canada. pp. 10–51.

[17] Cao, Z., Qin, T., Liu, T.Y., Tsai, M.F., Li, H., 2007. Learning to Rank: from Pairwise Approach to Listwise Approach, in: Proc. of the 24th International Conference on Machine Learning (ICML), Corvalis, Oregon, USA. pp. 129–136.

[18] Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., Federico, M., 2013. Report on the 10th IWSLT

Evaluation Campaign, in: Proc. of the 10th International Workshop on Spoken Language Translation (IWSLT), Heidelberg, Germany.

[19] Chieu, H.L., Ng, H.T., 2002. Named Entity Recognition: A Maximum Entropy Approach Using Global Information, in: Proc. of the 19th International Conference on Computational Linguistics (COLING), Association for Computational Linguistics, Taipei, Taiwan. pp. 1–7.

[20] Cho, E., Niehues, J., Ha, T.L., Sperber, M., Mediani, M., Waibel, A., 2016. Adaptation and Combination of NMT Systems: The KIT Translation Systems for IWSLT 2016, in: Proc. of the 13th International Workshop on Spoken Language Translation (IWSLT), Seattle, Washington, USA.

[21] Clémençon, S., Depecker, M., Vayatis, N., 2013. Ranking Forests. Machine Learning Research 14, 39–73.

[22] Evermann, G., Woodland, P., 2000a. Large Vocabulary Decoding and Confidence Estimation using Word Posterior Probabilities, in: Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Istanbul, Turkey. pp. 1655–1658.

[23] Evermann, G., Woodland, P., 2000b. Posterior Probability Decoding, Confidence Estimation and System Combination, in: Proc. of NIST Speech Transcription Workshop, College Park, MD, USA.

[24] Facco, A., Falavigna, D., Gretter, R., Vigano, V., 2006. Design and Evaluation of Acoustic and Language Models for Large Scale Telephone Services. Speech Communication 48, 176–190.

[25] Falavigna, D., Matassoni, M., Jalalvand, S., Negri, M., Turchi, M., 2016. DNN adaptation by automatic quality estimation of ASR hypotheses. Computer Speech and Language .

[26] Federico, M., Bentivogli, L., Paul, M., Stüker, S., 2012. Overview of the IWSLT 2012 Evaluation Campaign, in: Proc. of the 9th International Workshop on Spoken Language Translation (IWSLT), Hong Kong. pp. 11–27.

[27] Fiscus, J.G., 1997. A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER), in: Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE, Santa Barbara, CA, USA. pp. 347–354.

[28] Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian Network Classifiers. Machine Learning 29, 131–163.

[29] Fürnkranz, J., Hüllermeier, E., 2003. Pairwise Preference Learning and Ranking, in: Proc. of the 14th European Conference on Machine Learning (ECML), Springer, Cavtat, Croatia. pp. 145–156.

[30] Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely Randomized Trees. Machine learning 63, 3–42.

[31] Gillick, L., Cox, S.J., 1989. Some Statistical Issues in the Comparison of Speech Recognition Algorithms, in: Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Glasgow, Scotland. pp. 532–535.

[32] Goldwater, S., Jurafsky, D., Manning, C.D., 2010. Which Words Are Hard to Recognize? Prosodic, Lexical, and Disfluency Factors that Increase Speech Recognition Error Rates. Speech Communication 52, 181–200.

[33] Guerrero, C., Omologo, M., 2014. Exploiting Inter-microphone Agreement for Hypothesis Combination in Distant Speech Recognition, in: Proc. of the 22nd European Signal Processing Conference (EUSIPCO), Lisbon, Portugal. pp. 2385–2389.

[34] Hillard, D., Hoffmeister, B., Ostendorf, M., Schlueter, R., Ney, H., 2007. iROVER: Improving System Combination with Classification, in: Proc. of the North American Chapter of the Association for Computational Linguistics (NAACL), Rochester, New York. pp. 65–68.

[35] Hoffmeister, B., Klein, T., Schlüter, R., Ney, H., 2006. Frame Based System Combination and a Comparison with Weighted ROVER and CNC, in: Proc. of the 9th International Conference on Spoken Language Processing (INTERSPEECH), Pittsburgh, PA, USA. pp. 537–540.

[36] Hori, T., Chen, Z., Erdogan, H., Hershey, J., Le Roux, J., Mitra, V., Watanabe, S., 2015. The MERL/SRI system for the 3rd CHiME challenge using beamforming, robust feature extraction, and advanced speech recognition, in: Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE, Scottsdale, Arizona, USA.

[37] Jalalvand, S., Falavigna, D., 2015. Stacked Auto-Encoder for ASR Error Detection and Word Error Rate Prediction, in: Proc. of the 16th Annual Conference of the International Speech Communication

Association (INTERPSEECH), Dresden, Germany. pp. 2142–2146.

[38] Jalalvand, S., Falavigna, D., Matassoni, M., Svaizer, P., Omologo, M., 2015a. Boosted Acoustic Model Learning and Hypotheses Rescoring on the CHiME-3 Task, in: Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Scottsdale, Arizona, USA. pp. 409–415.

[39] Jalalvand, S., Negri, M., Falavigna, D., Turchi, M., 2015b. Driving ROVER with Segment-based ASR Quality Estimation, in: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL), Beijing, China. pp. 1095–1105.

[40] Knapp, C., Carter, G., 1976. The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech, and Signal Processing 24, 320–327.

[41] Kumatani, K., Arakawa, T., Yamamoto, K., McDonough, J., Raj, B., Singh, R., Tashev, I., 2012. Microphone Array Processing for Distant Speech Recognition: Towards Real-World Deployment, in: Proc. of Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), Hollywood, CA, USA. pp. 1–10.

[42] Kumatani, K., McDonough, J., Lehman, J.F., Raj, B., 2011. Channel Selection Based on Multichannel Cross-correlation Coefficients for Distant Speech Recognition, in: Proc. of the 3rd Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA), Edinburgh, UK. pp. 1–6.

[43] Lamel, L., Gauvain, J.L., Adda, G., 2001. Investigating Lightly Supervised Acoustic Model Training, in: Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Salt Lake City, USA. pp. 477–480.

[44] Li, X., Singh, R., Stern, R.M., 2002. Lattice Combination for Improved Speech Recognition, in: Proc. of the International Conference of Spoken Language Processing, Denver, CO, USA.

[45] Litman, D.J., Hirschberg, J.B., Swerts, M., 2000. Predicting Automatic Speech Recognition Performance Using Prosodic Cues, in: Proc. of the 1st North American chapter of the Association for Computational Linguistics Conference (NAACL), Seattle, Washington, USA. pp. 218–225.

[46] Loesch, B., Yang, B., 2010. Adaptive segmentation and separation of determined convolutive mixtures under dynamic conditions, in: International Conference on Latent Variable Analysis and Signal Separation, Springer. pp. 41–48.

[47] Mammone, A., Turchi, M., Cristianini, N., 2009. Support vector machines. Wiley Interdisciplinary Reviews: Computational Statistics 1, 283–289.

[48] Mangu, L., Brill, E., Stolcke, A., 2000. Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks. Computer Speech & Language 14, 373–400.

[49] McFee, B., Lanckriet, G.R., 2010. Metric Learning to Rank, in: Proc. of the 27th International Conference on Machine Learning (ICML), Haifa, Israel. pp. 775–782.

[50] Mestre, X., Lagunas, M., 2003. On Diagonal Loading for Minimum Variance Beamformers, in: Proc. of the 3rd IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Darmstadt, Germany. pp. 459–462.

[51] Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., Khudanpur, S., 2010. Recurrent Neural Network Based Language Model, in: Proc. of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH), Makuhari, Chiba, Japan. pp. 1045–1048.

[52] Mowlaee, P., Saeidi, R., 2013. Iterative closed-loop phase-aware single-channel speech enhancement. IEEE Signal Processing Letters 20, 1235–1239.

[53] Negri, M., Turchi, M., C. de Souza, J.G., Daniele, F., 2014. Quality Estimation for Automatic Speech Recognition, in: Proc. of the 25th International Conference on Computational Linguistics: Technical Papers (COLING), Dublin, Ireland. pp. 1813–1823.

[54] Paul, D.B., Baker, J.M., 1992. The Design for the Wall Street Journal-based CSR Corpus, in: Proc. of the Workshop on Speech and Natural Language, Stroudsburg, PA, USA. pp. 357–362.

[55] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine Learning in Python. Machine Learning Research 12, 2825–2830.

[56] Pellegrini, T., Trancoso, I., 2010. Improving ASR Error Detection with Non-decoder Based Features, in: Proc. of the 11th Annual Conference of the International Speech Communication Association (IN-

TERSPEECH), Makuhari, Chiba, Japan. pp. 1950–1953.

[57] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 2011. The Kaldi Speech Recognition Toolkit, in: Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Hawaii, USA.

[58] Riccardi, G., Hakkani-Tur, D., 2005. Active Learning: Theory and Applications to Automatic Speech Recognition. IEEE Transaction on Speech and Audio Processing 13, 504–511.

[59] Schwenk, H., Gauvain, J.L., 2000. Improved ROVER using Language Model Information, in: Proc. of ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW), Orsay, France. pp. 47–52.

[60] Soricut, R., Echihabi, A., 2010. TrustRank: Inducing Trust in Automatic Translations via Ranking, in: Proc. of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, Uppsala, Sweden. pp. 612–621.

[61] C. de Souza, J.G., Buck, C., Turchi, M., Negri, M., 2013. FBK-UEdin participation to the WMT13 Quality Estimation shared-task, in: Proceedings of the Eighth Workshop on Statistical Machine Translation, pp. 352–358.

[62] C. de Souza, J.G., González-Rubio, J., Buck, C., Turchi, M., Negri, M., 2014. FBK-UPV-UEdin participation in the WMT14 Quality Estimation shared-task, in: Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, MD, USA.

[63] C. de Souza, J.G., Zamani, H., Negri, M., Turchi, M., Falavigna, D., 2015. Multitask Learning for Adaptive Quality Estimation of Automatically Transcribed Utterances, in: Proc. of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT), Denver, Colorado, USA. pp. 714–724.

[64] Specia, L., Cancedda, N., Dymetman, M., Turchi, M., Cristianini, N., 2009. Estimating the Sentence-Level Quality of Machine Translation Systems, in: Proc. of the 13th Annual Conference of the European Association for Machine Translation (EAMT), Barcelona, Spain. pp. 28–35.

[65] Stolcke, A., 2011. Making the Most From Multiple Microphones in Meeting Recognition, in: Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic. pp. 4992–4995.

[66] Tam, Y.C., Lei, Y., Zheng, J., Wang, W., 2014. ASR Error Detection using Recurrent Neural Network Language Model and Complementary ASR, in: Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Florence, Italy. pp. 2312–2316.

[67] Turchi, M., Negri, M., Federico, M., 2013. Coping with the subjectivity of human judgements in MT quality estimation, in: Eighth Workshop on Statistical Machine Translation (WMT), pp. 240–251.

[68] Ueffing, N., Ney, H., 2007. Word-Level Confidence Estimation for Machine Translation. Computational Linguistics 33, 9–40.

[69] Vacher, M., Lecouteux, B., Portet, F., 2012. Recognition of voice commands by multisource asr and noise cancellation in a smart home environment, in: 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), pp. 1663–1667.

[70] Vesely, K., Ghoshal, A., Burget, L., Povey, D., 2011. Sequence-discriminative Training of Deep Neural Networks, in: Proc. of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH), Florence, Italy. pp. 2345–2349.

[71] Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., Matassoni, M., 2013. The Second CHiME speech Separation and Recognition Challenge: An Overview of Challenge Systems and Outcomes, in: Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Olomouc, Czech Republic. pp. 162–167.

[72] Wang, H., Ragni, A., Gales, M.J., Knill, K.M., Woodland, P.C., Zhang, C., 2015. Joint decoding of tandem and hybrid systems for improved keyword spotting on low resource languages, in: Proceedings of Interspeech.

[73] Wessel, F., Schluter, R., Macherey, K., Ney, H., 2001. Confidence Measures for Large Vocabulary Continuous Speech Recognition. IEEE Transactions on Audio, Speech and Language Processing 9,

288–298.

[74] Wölfel, M., Fgen, C., Ikbal, S., McDonough, J.W., 2006. Multi-Source Far-Distance Microphone Selection and Combination for Automatic Transcription of Lectures, in: Proc. of the 9th International Conference on Spoken Language Processing (INTERSPEECH), Pittsburgh, Pennsylvania. pp. 361–364.

[75] Wölfel, M., McDonough, J., 2005. Combining Multi-Source Far Distance Speech Recognition Strategies, in: Proc. of the 9th European Conference on Speech Communication and Technology (EUROSPEECH), Lisbon, Portugal. pp. 3149–3152.

[76] Wölfel, M., McDonough, J., 2009. Distant Speech Recognition. John Wiley & Sons Ltd.

[77] Yang, J., Zhang, C., Ragni, A., Gales, M.J.F., Woodland, P.C., 2016. System combination with log-linear models, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5675–5679.

[78] Zhang, R., Rudnicky, A.I., 2006. Investigations of Issues for Using Multiple Acoustic Models to Improve Continuous Speech Recognition, in: Proc. of the 9th International Conference on Spoken Language Processing (INTERSPEECH), Pittsburgh, PA, USA.

[79] Zhou, K., Xue, G.R., Zha, H., Yu, Y., 2008. Learning to Rank with Ties, in: Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA. pp. 275–282.