Scaling Acoustic and Language Model Probabilities in a CSR System[⋆]

Amparo Varona[1] and M. Inés Torres[1]

Departamento de Electricidad y Electrónica, Fac. de Ciencia y Tecnología,
UPV/EHU, Apartado 644, 48080 Bilbao, Spain
{amparo,manes}@we.lc.ehu.es

Abstract. It is well known that a direct integration of acoustic and language models (LM) into a Continuous Speech Recognition (CSR) system leads to low performances. This problem has been analyzed in this work as a practical numerical problem. There are two ways to get optimum system performances: scaling acoustic or language model probabilities. Both approaches have been analyzed from a numerical point of view. They have also been experimentally tested on a CSR system over two Spanish databases. These experiments show similar reductions in word recognition rates but very different computational cost behaviors. They also show that the values of scaling factors required to get optimum CSR systems performances are closely related to other heuristic parameters in the system like the beam search width.

1 Introduction

Integration of language and acoustic models in a Continuous Speech Recognition (CSR) system is invariably based on the well-known Bayes' rule, i.e., the recognizer must find the word sequence $\hat{\Omega}$ that satisfies:

$$\hat{\Omega} = \arg\max_{\Omega} P(\Omega)P(A/\Omega) \tag{1}$$

where $P(\Omega)$ is the probability that the word sequence $\Omega \equiv \omega_1\omega_2\ldots\omega_{|\Omega|}$ from some previously established finite vocabulary $\Sigma = \{\omega_j\}$, $j = 1\ldots|\Sigma|$, will be uttered and $P(A/\Omega)$ is the probability of the sequence of acoustic observations $A = a_1a_2...a_{|A|}$ for a given sequence of words Ω. Probabilities $P(A/\Omega)$ are represented by acoustic models, usually Hidden Markov Models (HMM). The *a priori* probabilities $P(\Omega)$ are given by the Language Model (LM).

However, the combination of acoustic and LM probabilities obtained through Equation 1 usually leads to poor CSR system performances. In fact, it is well known that best performances of a CSR system are obtained when acoustic and language model probabilities in the Bayes' rule are modified by introducing exponential scaling factors [1] [2] [3].

From a theoretical point of view, the scaling parameters are needed because acoustic and LM probability distributions are not real but approximations [1]. The two probability distributions are estimated independently using different stochastic models that model different knowledge sources. Moreover, the parameters of the acoustic and language models are estimated on the basis of speech and text data corpora, respectively. Thus, scaling parameters need to be applied to lessen these effects and then obtain good system performances.

In practice a numerical problem needs to be solved: the acoustic probabilities are normally much smaller than those of the LM. In addition, they appear more times. As a consequence, the contribution of LM probabilities could not be relevant [4] to obtain the most probable word sequence $\hat{\Omega}$. Therefore, low CSR system performances are usually obtained.

Section 2 illustrates this problem through an example of word sequence recognition. The usual way to get better system performances is to reduce the LM probabilities using a scaling factor $\alpha > 1$: $P(\Omega)^\alpha$. Alternatively, the acoustic probabilities can be increased by using a scaling factor $\gamma < 1$: $P(\Omega/A)^\gamma$. The effects of both scaling procedures are also analyzed in Section 2. Section 3 shows an experimental evaluation of both scaling procedures. The CSR system performance was measured in terms of both the Word Error Rate (WER) and the involved computational cost, which are also related with the heuristic beam-search applied. The experiments were carried out on two Spanish application tasks including read and spontaneous speech respectively. Therefore these databases represent two different levels of difficulty. Finally, some concluding remarks are given in Section 4.

2   Scaling Factors over Acoustic and LM Probability Distributions

Most of the current CSR systems use the one-pass Viterbi algorithm to obtain a sequence of decoded words given a sequence of acoustic observations according to Equation 1. The Viterbi-based search solves a maximization problem at each time t by choosing the higher accumulated probability for each node of the trellis. The argument of the maximization, i.e. the node at time $t-1$ leading to maximum accumulated probability at time t, is also saved. Thus, finally the algorithm recovers in a backward step the most probable path of trellis nodes according to the whole acoustic observations sequence and, as a consequence, the most probable sequence of words uttered. Figure 1 shows this procedure for a small vocabulary of five words: $w_1, ... w_5$ at time t and time $t+1$. $-log P_{w_i}$ [1] represents accumulated probabilities corresponding to partial paths ending at each word w_i, i.e. at trellis nodes that match the final state of word w_i at time t. For these nodes, the LM has to be checked to get the $-log P(w_i/w_j)$.

---

[1] In practice probabilities are managed as minus their logarithm to avoid numbers quickly falling to zero. We keep this representation in the example to better analyze differences among probability values. Thus lower values stand for higher probable paths.

Then, the accumulated probability $-logP_{w_i}$ at time $t+1$ is obtained through a new maximization procedure that includes this time combinations with LM probabilities. As a consequence of this maximization, a new word is added to the partial path under consideration.
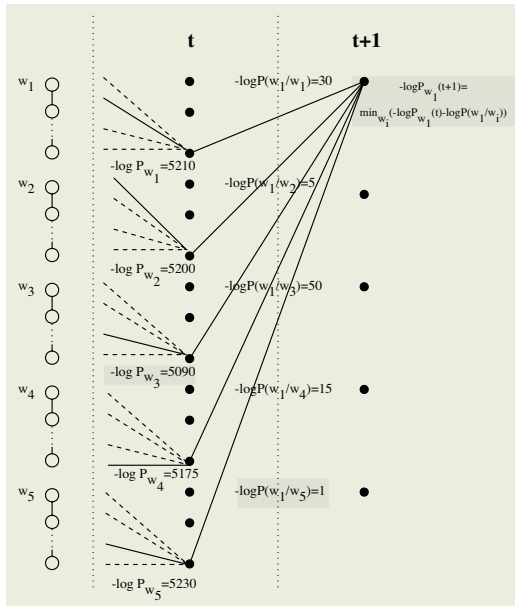

Fig. 1. The Viterbi based search choose at each time t the highest accumulated probability for each node of the trellis. $-logP_{w_i}$ represents accumulated probabilities for nodes matching the final state of a word model at time t. For these nodes, the language model has to be checked to get the $-logP(w_i/w_j)$. Then, the accumulated probability $-logP_{w_i}$ at time $t+1$ is obtained through a new maximization procedure that includes this time combinations with language model probabilities.

Figure 1 also shows the numerical problem previously discussed. The example in Figure 1 shows that the most probable partial path at time t ends at word w_3 and the less probable one ends at word w_5. According to the bigram model probabilities $P(w_1/w_j)$ the most probable word to be added at time t to the partial path is word w_5 and the less probable is w_3, when the word w_1 is considered at time $t+1$. Then, a new maximization has to be solved to estimate the accumulated probabilities at time $t+1$. For trellis node matching the initial state of word w_1:

$$P_{w_1}^{t+1} = \max_{w_i}(P_{w_i}^t P(w_1/w_i)) \rightarrow -log(P_{w_1}^{t+1}) = \min_{w_i}(-logP_{w_i}^t - logP(w_1/w_i)) \quad (2)$$

Table 1 shows the accumulated probabilities $-logP_{w_1/w_i}^{t+1}$ at time $t+1$ and initial state of word w_1 for all possible recognized word w_i at time t for the

Table 1. Accumulated probabilities $-log P^{t+1}_{w_1/w_i}$ at time $t+1$ and initial state of word w_1 for all possible recognized word w_i at time t for the example in Figure 1.

$-log P^{t+1}_{w_1/w_1}$	$-log P^{t+1}_{w_1/w_2}$	$-log P^{t+1}_{w_1/w_3}$	$-log P^{t+1}_{w_1/w_4}$	$-log P^{t+1}_{w_1/w_5}$
5249	5205	**5140**	5190	5231

example in Figure 1. Maximization in 2 is solved by choosing among values in Table 1. Thus word w_3 is selected to be added to this partial path, even if it is the less probable word according to the LM probabilities.

This example illustrates the numerical problem when Equation 1 is directly applied to decodeingan utterance without any scaling probabilities. Accumulated probabilities at the end of each partial sequence of words in the Viterbi trellis mainly depend on the values of acoustic $P(A/\Omega)$ probabilities because they are usually much smaller than those of the LM $P(\Omega)$. Moreover, acoustic probabilities appear, i.e. should be multiplied (or be summed their minus logarithm), at each node of the trellis whereas LM probabilities are only considered for trellis nodes matching final states of each word in the task lexicon. The gap among accumulated probabilities is therefore usually bigger than the gap among LM probabilities (see Figure 1). The immediate consequence is that LM probabilities are irrelevant in most situations when choosing the best, most probable, partial path [4]. The scarce contribution of the LM probabilities leads to low CSR system performances.

There are two easy ways to avoid this situation: to increase the gap among LM probabilities or to decrease the gap among accumulated probabilities. In the first case the LM probabilities are scaled by an exponential parameter $\alpha > 1$:$(P(\omega))^\alpha$. In the second one the acoustic model probabilities are scaled by $\gamma < 1$:$(P(A/\omega))^\gamma$ When LM probabilities are raised to a power $\alpha > 1$: $(P(\omega))^\alpha$, all of them are attenuated, but this attenuation is higher for lower probability values. The gap between high and low probability values is now bigger. The LM probability values become more competitive with the increase of α values, up to a maximum where LM probabilities are overvalued. As a consequence, LM probabilities are now more relevant to choose the best partial path. Table 2 shows the accumulated probabilities at time $t+1$ for the example in Figure 1 and Table 1 when LM probabilities have been scaled by different values of parameter α. This Table shows that word w_2 ($\alpha = 4$ and $\alpha = 6$) or w_4 ($\alpha = 8$) could now be selected to be added to partial path, even if they correspond to partial paths with low accumulated probabilities (see Figure 1). When acoustic probabilities are raised to a power $\gamma < 1$: $(P(A/\omega))^\gamma$ all of them are increased but the increase is higher for lower probability values. The gap among accumulated probabilities is reduced in this case because accumulated probability values mainly depend on acoustic probabilities. The LM probabilities are also more and more competitive when γ decreases down to a minimum. As a consequence, LM probabilities are now also more relevant to choose the best partial path. Table 3 shows the accumulated probabilities at time $t+1$ for example in Figure 1 and Table 1 when acoustic probabilities have been scaled by different values of parameter γ.

This Table shows that word w_2 ($\gamma = 0.4$), w_3 ($\gamma = 0.5$) or w_4 ($\gamma = 0.3$) could be now selected to be added to partial path, even if they correspond to partial paths with low accumulated probabilities at time t (see Figure 1).

Table 2. Accumulated probabilities $-log P_{w_1/w_i}^{t+1}$ at time $t+1$ and initial state of word w_1 for all possible recognized word w_i at time t for the example in Figure 1 and Table 1. LM probabilities have been scaled by $\alpha = 4, 6, 8$.

α	$-log P_{w_1/w_1}^{t+1}$	$-log P_{w_1/w_2}^{t+1}$	$-log P_{w_1/w_3}^{t+1}$	$-log P_{w_1/w_4}^{t+1}$	$-log P_{w_1/w_5}^{t+1}$
4	5330	**5220**	5340	5235	5234
6	5390	**5230**	5440	5265	5236
8	5450	5240	5540	5295	**5238**

Table 3. Accumulated probabilities $-log P_{w_1/w_i}^{t+1}$ at time $t+1$ and initial state of word w_1 for all possible recognized word w_i at time t for the example in Figure 1 and Table 1. Acoustic probabilities have been scaled by $\gamma = 0.5, 0.4, 0.3$.

γ	$-log P_{w_1/w_1}^{t+1}$	$-log P_{w_1/w_2}^{t+1}$	$-log P_{w_1/w_3}^{t+1}$	$-log P_{w_1/w_4}^{t+1}$	$-log P_{w_1/w_5}^{t+1}$
0.5	2635	2605	**2595**	2602	2616
0.4	2114	**2085**	2086	**2085**	2093
0.3	1593	1565	1577	**1555**	1570

When no scaling factors are applied ($\alpha = 1$ and $\gamma = 1$), the most probable partial path includes word w_3 (see Table 1). However, for $\alpha > 1$ (Table 2) and for $\gamma < 1$ (see Table 3) the most probable partial paths include word w_2. High values of α and low values of γ are not usually considered since LM probabilities are overvalued leading to low recognition rates.

CSR systems introduce a beam-search in order to reduce the computational cost of the search. The maximum probability of all trellis nodes, i.e. of all partial path is calculated at each time t, $1 \leq t \leq |A|$. Then, partial paths with accumulated probabilities under a percentage of this maximum are not considered any more. The value of an heuristic parameter $0 \leq bf \leq 1$ controls the number of partial paths considered. This parameter should be adjusted to reduce the computational cost without reducing the recognition rate. The number of partial paths to be evaluated mainly depends on the gap among accumulated probabilities corresponding to different partial paths at each time t. Both scaling factors modify partial path probabilities. However, scaling LM probabilities ($\alpha > 1$) increases the gap among accumulated probabilities whereas scaling acoustic probabilities ($\gamma < 1$) reduces it. As a consequence, the number of partial paths considered, and the involved computational cost, are expected to be very different for the two scaling procedures: higher for γ scaling than for α one.

3   Experimental Evaluation

The experimental evaluation was carried out on two Spanish databases with very different levels of difficulty: Bdgeo and Info_Tren.

Bdgeo is a task-oriented Spanish speech corpus [5] consisting of 82000 words and a vocabulary of 1208 words. This corpus represents a set of queries to a Spanish geography database. This is a specific task designed to test integrated systems for automatic speech understanding. Acoustic models were previously trained over a phonetic database consisting of 1530 sentences uttered by 47 speakers. The ML training corpus consisted of 9150 sentences. The test set consisted of 600 sentences uttered by 12 speakers in a laboratory environment at 16KHz. Perplexity values for this task were 13.1 for n=2, 7,53 for n=3 and 7,17 for n=4.

Info_Tren database was recorded as part of a Spanish project to develop a dialogue system. The task consisted of 227 Spanish dialogues on train information resulting in a vocabulary of around 2000 words. Info_tren is a difficult task since it was recorded through the telephone (8Khz), applying the well known *Wizard of Oz* mechanism. Thus, it is a spontaneous speech database were many different types of disfluencies are found: speaker and background noises, filled pauses, lengthenings, etc. [6]. Info_tren is the first spontaneous dialogue database recorded by Castilian Spanish speakers. The training corpus consisted of 191 dialogues uttered by 63 different speakers (1349 user turns resulting in 16500 words plus 5000 disfluencies). The test set consisted of 36 new dialogues uttered by 12 new speakers (308 user turns including 4000 words plus around 500 disfluencies). Perplexity values for this task were 36,8 for n=2, 34,8 for n=3 and 36,3 for n=4.

Uttered sentences were decoded by the time-synchronous Viterbi algorithm. In order to reduce the computational cost a beam-search algorithm was applied with different widths: high values of bf parameter for narrow beams and low values of bf for wide beams. The experiments were carried out using standard back-off n-gram models. Two series of experiments were carried out. In the first series LM probabilities were scaled by $|\alpha > 1 \ (P(\omega))^{\alpha}$. Table 4 shows the CSR system performance measured in terms of both, the Word Error Rate (WER) and the Average number of Active Nodes (AAN) in the lattice, including both acoustic and language model states. It shows WER and ANN for $\alpha = 1 \ldots 7$ and $bf = 0.6, 0.5, 0.4$ values for three different n-gram ($n = 2, 3, 4$) models. Optimum performances are emphatized and underlined.

These experiments confirm that the scaling Language Model probabilities clearly leads to better CSR systems performances. Table 4 shows that both the word error rate and the average number of nodes in the lattice increased with α (up to a minimum), for any n-gram model and value of bf parameter. This Table also shows a relationship between the bf and α values required to get optimum performances: higher values of α are needed for wider beams search.

In the second series of experiments the acoustic probabilities have been scaled by $\gamma < 1 \ (P(A/\omega))^{\gamma}$. Table 5 shows the CSR system performances obtained through these experiments: the word error rate and the average number of active nodes in the lattice for three different n-gram ($n = 2, 3, 4$) models and different values of bf and γ parameters. Optimum performances are also emphatized and underlined.

These experiments confirm that the scaling acoustic model probabilities clearly leads to better CSR systems performances measured in terms of word

Table 4. CSR performance obtained using n-grams LMs with $n = 2 \ldots 4$ for Bdgeo task and Info_tren task. Different values of the scaling factor over LM probabilities (α) and beam-search (bf) were applied.

α	BDGEO						Info_tren					
	bf=0.6		bf=0.5		bf=0.4		bf=0.6		bf=0.5		bf=0.4	
	WER	AAN	WER	AAN	WER	AAN	WER	AAN	WER	AAN	WER	AAN
n=2												
1	41.89	1824	41.62	3964	41.87	6358	62.15	1363	61.69	3260	62.16	4798
2	26.43	781	25.8	2588	25.93	4843	51.12	751	50.23	2594	51.14	4038
3	21.65	280	20.22	1508	20.34	3456	**47.06**	**369**	43.83	1912	45.52	3217
4	**20.54**	**114**	16.99	764	16.72	2330	47.91	190	41.08	1291	43.95	2422
5	24.76	62	**15.80**	**380**	15.33	1441	51.50	97	**39.60**	**799**	42.56	1734
6	31.42	40	15.95	218	14.81	858	59.37	54	40.32	467	**42.29**	**1204**
7	40.39	28	17.01	143	**14.29**	**526**	66.61	36	41.75	294	42.34	855
n=3												
1	44.18	2027	38.85	5189	38.85	9737	62.02	2011	58.69	6400	59.10	9998
2	22.43	746	21.86	2984	21.85	6659	51.30	1035	48.72	4668	53.08	5213
3	**17.19**	**238**	15.35	1529	15.18	4207	**47.31**	**488**	42.14	3172	47.10	3954
4	17.51	90	11.74	702	11.6	2526	47.68	240	38.72	1978	44.68	3112
5	21.79	48	10.85	328	10.32	1426	52.48	118	**38.01**	**1135**	42.73	2437
6	29.52	32	**10.82**	**179**	9.65	800	60.04	66	38.41	631	**41.13**	**1935**
7	36.81	23	13.04	114	**9.45**	**467**	68.43	41	41.58	378	43.01	968
n=4												
1	38.75	2052	38.50	5374	38.51	9930	62.10	2122	58.80	6480	60.01	10112
2	21.77	746	21.86	3053	21.13	7041	51.40	1193	48.90	4720	53.10	5293
3	**16.41**	**235**	14.44	1544	14.35	4379	**47.48**	**593**	42.25	3286	47.30	4015
4	16.71	89	10.92	704	10.82	2593	47.96	310	38.83	2229	44.78	3210
5	21.11	47	10.24	328	9.45	1451	52.56	195	**37.84**	**1269**	42.96	2563
6	28.64	31	**10.22**	**177**	8.72	808	60.34	92	38.63	702	**41.20**	**2078**
7	35.84	23	12.48	113	**8.83**	**469**	67.13	68	42.31	415	43.60	1050

error rates. However, the reduction of word error rates is obtained along with important raises of the average number of active nodes in the lattice. Moreover, in this case a wide beam-search keeps a very high number of active hypotheses in the lattice increasing not only the involved computational cost but also the number of recognition errors. Table 5 also shows a relationship between the bf and γ values required to get optimum performances: lower values of γ are needed for wider beam search.

4   Concluding Remarks

In this work, we have experimentally analyzed the bad relationship between acoustic and LM probabilities when they are directly integrated using de Bayes' rule in CSR systems. A practical numerical problem arises since acoustic probabilities are much smaller than language ones. In addition they appear more times. As consequence, the contribution of LM probabilities could not be relevant when choosing the best path in the trellis. The gap among LM probabilities can be increased by scaling them by an exponential factor $\alpha > 1$ or the gap among accumulated probabilities can be reduced by scaling acoustic probabilities by an exponential factor $\gamma < 1$. Both possibilities have been analyzed and tested on a CSR system over two Spanish databases. Both scaling approaches clearly led to important reductions of word error rates. However, α scaling of LM probabilities

Table 5. CSR performance obtained using n-grams LMs with $n = 2 \ldots 4$ for Bdgeo and Info_tren task. Different values of the scaling factor over acoustic probabilities (γ) and beam-search (bf) were applied.

γ	BDGEO						Info_tren					
	bf=0.9		bf=0.8		bf=0.6		bf=0.8		bf=0.6		bf=0.5	
	WER	AAN	WER	AAN	WER	AAN	WER	AAN	WER	AAN	WER	AAN
						n=2						
1	68.58	20	46.91	243	41.89	1824	93.53	178	61.15	1363	61.69	3260
0.8	61.31	24	39.50	341	35.98	2484	92.06	350	58.03	2133	58.06	4334
0.5	44.34	39	27.15	708	25.60	4620	**51.45**	**793**	**43.92**	**4051**	**50.87**	**7247**
0.2	19.63	123	**15.13**	**2775**	**24.87**	**6538**	67.21	2987	67.21	9369	67.42	11610
0.1	**19.15**	**358**	17.45	7248	26.10	9361	91.48	6055	91.81	11455	91.99	12336
0.05	46.54	1076										
						n=3						
1	65.88	18	44.58	222	44.18	2027	93.82	204	60.02	2011	58.69	9980
0.8	60.34	20	37.24	315	33.12	2905	92.13	416	56.28	3364	**56.30**	**12415**
0.5	42.23	30	22.49	668	22.47	6402	**50.03**	**1024**	**41.77**	**5830**	59.77	18780
0.2	16.12	94	**10.4**	**3172**	**17.34**	**9000**	64.86	4010	53.60	12360		
0.1	**12.04**	**303**	10.95	12031	21.10	12314	80.04	8135				
0.05	44.85	1424										
						n=4						
1	65.23	18	44.36	221	38.75	2052	93.90	230	62.10	2122	58.80	6480
0.8	59.82	20	36.75	313	32.75	2961	92.23	426	55,35	3650	**54.30**	**13415**
0.5	42.12	30	21.89	668	21.26	6757	**49.92**	**1068**	**40.13**	**6930**	63.10	19565
0.2	16.63	93	**9.67**	**3313**	**17.10**	**9990**	63.10	4340	52.15	13360		
0.1	**11.74**	**308**	10.85	13730	20.90	12816	79.10	8240				
0.05	43.12	1472										

also led to important reductions of the computational costs whereas γ scaling of acoustic probabilities led to undesirable raises. It has also been shown that these effects are not independent of other heuristic phenomena in the system like the beam-search width. Thus, the values of scaling factors in CSR systems should be experimentally established along with other heuristic parameters.

References

1. Jelinek, F.: Five speculations (and a divertimento) on the themes of h. bourlard, h. hermansky and n. morgan. Speech Communication **18** (1996) 242–246
2. Rubio, J.A., Diaz-Verdejo, J.E., García, P., Segura, J.C.: On the influence of of frame-asynchronous grammar scoring in a csr system. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing. Volume II. (1997) 895–899
3. A.Ogawa, Takeda, K., Itakura, F.: Balancing acoustic and linguistic probabilities. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing. Volume II. (1998) 181–185
4. Varona, A., Torres, I.: High and low smoothed lms in a csr system. Progress in Pattern Recognition Speech and Image Analysis. Computer. LNCS **1** (2003) 236–243
5. Díaz, J., Rubio, A., Peinado, A., Segarra, E., Prieto, N., F.Casacuberta: Albayzin: a task-oriented spanish speech corpus. In: First Int. Conf. on language resources and evaluation. Volume II. (1998) 497–501
6. Rodríguez, L., Torres, I., Varona, A.: Evaluation of sublexical and lexical models of acoustic disfluencies for spontaneous speech recognition in spanish. In: Proc. of European Conference on Speech Technology. Volume 3. (2001) 1665–1668