Automatic Detection of Well Recognized Words in Automatic Speech Transcriptions

Julie Mauclair, Yannick Estève, Simon Petit-Renaud, Paul Deléglise

Laboratoire d'Informatique de l'Université du Maine
Le Mans, France
{mauclair,esteve,petit-renaud,deleglise}@lium.univ-lemans.fr

Abstract

This work adresses the use of confidence measures for extracting well recognized words with very low error rate from automatically transcribed segments in a unsupervised way. We present and compare several confidence measures and propose a method to merge them into a new one. We study its capabilities on extracting correct recognized word-segments compared to the amount of rejected words. We apply this fusion measure to select audio segments composed of words with a high confidence score. These segments come from an automatic transcription of french broadcast news given by our speech recognition system based on the CMU Sphinx3.3 decoder. Injecting new data resulting from unsupervised treatments of raw audio recordings in the training corpus of acoustic models gives statistically significant improvement (95% confident interval) in terms of word error rate. Experiments have been carried out on the corpus used during ESTER, the french evaluation campaign.

1. Introduction

Confidence measures are used on various applications of speech processing like speech recognition(Wessel and Ney, 2005; Cox and Dasmahapatra, 2002), dialog (San-Segundo et al., 2001) and language identification (Metze et al., 2000). They help to decide if an hypothesis is right or wrong.

Moreover, to train models for a new recognition system, we need large amounts of speech data. Nowadays, large collections of speech data are available but unfortunately, most of them are without transcriptions and has to be transcribed manually. Manual transcription of audio recordings is high-cost, which limits the size of the training corpora for the models. Performances of a speech recognition system rise mainly from the quality of the acoustic models. These stastistical models are more robust as their training corpora are important and close to the application task. A low-cost method to increase the size of the training corpora is to add automatic transcriptions (Deléglise et al., 2005) but this method can introduce 'noise' in modeling due to errors in the recognizer transcriptions. For filtering those transcriptions, we can use closed captions (Chen et al., 2004). But this method imposes additional information about transcriptions. To avoid this problem, we can use confidence measures to select training hypothesis. In (Wessel and Ney, 2005), the authors use confidence measure estimated with posterior probabilities given by the recognizer. A disadvantage of word posteriors is the strong sensitivity of this measurement to the topology of the search space on which it is computed. This topology is affected by heurisitcs used during the search space generation to reduce its size and make the recognition possible on a reasonable time. Moreover, non negligible time processing is needed to compute word posteriors.

We propose in this article a preliminary study for automatic detection of well recognized words with confidence measures which are easilly computable and not affected by the search space size before testing word posteriors as confidence measure for the same task. Those measures come from two parts of the recognizer in order to merge them usefully. While the first measure is based on acoustic likelihood, the other one is computed from the language model thanks to the observation of the back-off behavior. We develop a fusion method to use a single metric for filtering. We evaluate the measures quality (fusion and single ones) on french broadcast news by using the Normalised Cross Entropy (NCE). We compare them by testing their ability to minimize the WER at constant rejection rate. We use the best fusion measure to choose audio segments which are automatically transcribed by our recognizer. The add of these data resulting from automatic treatments in the training corpus of acoustic models is used to improve our system performances. Finally, we show a new measure improving word posteriors taken as a confidence measure in terms of NCE which let us believe further future improvements for filtering data.

2. Confidence measures

Let a set composed of N recognized words $\{w_1, \ldots, w_N\}$. Each word w is associated with a confidence measure $m(w)$ following suitable properties: the measure should be in the usual domain $[0, 1]$ and the measure should be interpretable as a probability that the word w is correct. In consequence of the last property, we have: $\frac{1}{N}\sum_{i=1}^{N} m(w_i) \approx$ CWRR, where CWRR (Correct Word Retained Rate) is the correct recognition rate on the emitted words (deletions are not counted).

2.1. Acoustic confidence measure

This measure is based on the comparison of the acoustic likelihood provided by the speech recognition system for a given hypothesis to the one that would be provided by an unconstrained phone loop model (Cox and Dasmahapatra, 2002):

$$m^*_{ac}(w) = \frac{1}{N_f(w)}[\log P(Y|\lambda_C) - \log P(Y|\lambda_L)] \quad (1)$$

where w is the recognized word with N_f frames, Y is the sequence of acoustic observations, $P(Y|\lambda_C)$ is the acoustic score given by the recognizer model, and $P(Y|\lambda_L)$ is the acoustic score given by an unconstrained phone loop.
The measure proposed above does not belong to $[0, 1]$. Thus, we propose to add a new normalization using the sigmoid-like transformation presented at equation 2:

$$m_{ac}(w) = \frac{\exp\left(\frac{m_{ac}(w)-\mu}{\sigma}\right) + a}{\exp\left(\frac{m_{ac}(w)-\mu}{\sigma}\right) + 1} \quad (2)$$

where μ, σ are the average and the standard deviation of the initial acoustic measure and $a = 2\text{CWRR} - 1$: the second property approximation (see the begining of the section) with the development corpus.

2.2. Language confidence measure

Usually, a speech recognition system combines scores provided by acoustic models with probabilities given by a n-gram language model. Section 2.1. presented a confidence measure to evaluate the relevance of the acoustic models. It seems interesting to have an equivalent measure for the language model. In this part, we introduce a new measure designed for back-off n-gram language models.
In the linguistic confidence measure we introduce, we propose to use as information the LM back-off behavior (LMBB) (Uhrick and Ward, 1997): a given word recognized with a given left context is associated with the highest order of n-grams seen in the training corpus with this word and this context. For example, if the sequence of words 'it is the ninth time' is recognized using a quadrigram model and if the quadrigram [is the ninth time] was observed in the training corpus, 'time' will be associated with the order 4. But if this quadrigram was not observed, whereas the trigram [the ninth time] was, 'time' will be associated with the order 3. This is recurrent down to order 1 (or 0 if out-of-vocabulary words can be processed).
Moreover, it is known that an error occurring on a word has an impact on the correctness of the words located in the immediate context of this erroneous word. According to that, and assuming that the LM back-off behavior is an acceptable criterion to predict the correctness of a word, our measure considers the LM back-off behavior of left and right neighbors of a word in addition to the highest order of n-grams this word is associated with.
So, each word of a recognized hypothesis is associated with three values: the highest order of observed n-grams which the left neighbor word and its context are associated with, the corresponding word order itself and its context, and the corresponding order of the right neighbor word and its context.
These triplets can be used as classes. In order to reduce the number of classes, each word of a recognized hypothesis is finally associated with a three components label:

1. the symbol -, =, or + when the highest order of n-gram associated with its left neighbor word is respectively lower than, equal to, or higher than the highest order of n-gram associated with the considered word,
2. the highest order of observed n-grams which the word and its context are associated with
3. the symbol -, =, or + when the order of the right neighbor word is respectively lower than, equal to, or higher than the one of the considered word.

This label corresponds to a class of recognized words, decoded in the same context in terms of LM behavior.
By comparing a set of automatic transcriptions with words labeled with these triplets, with a manual transcription of the same set of sentences, we compute the error rate of each class of words. The error rate is the ratio of the number of misrecognized words (substitutions or insertions) included in this class *vs.* the number of recognized words in this class. The estimated error rate for a given class will later be used as the confidence measure for words of this class during processing of test data. This measure will be called the LMBB confidence measure, and the value of this measure for a word w will be noted $m_{lmbb}(w)$.
Figure 1 shows that there is a correlation between the LM back-off behavior around a recognized word and the WER.

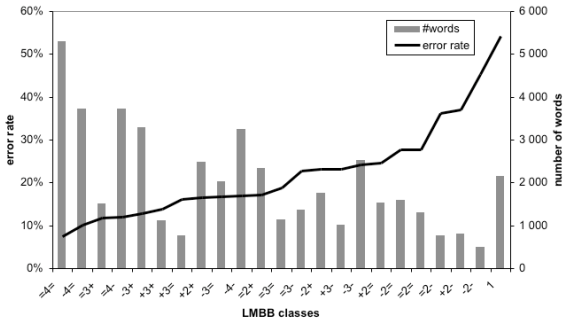

Figure 1: Error rate and word distribution per LM back-off behavior (LMBB) class on the training data

3. Corpus and system training

Experiments have been carried out on the ESTER corpus. ESTER is an evaluation campaign of french broadcast news transcriptions systems which started in 2003 and completed in January 2005 (Galliano et al., 2005). The system used there by the Laboratoire d'Informatique de l'Université du Maine (LIUM) is based on the CMU Sphinx 3.3 decoder. The data were recorded from six radios: *France Inter, France Info, RFI, RTM, France Culture* and *Radio Classique*. The data are divided into three sets; only the two first ones are annotated[1]. Shows (10 minutes up to 60 minutes) from those two first sets contain few silence, music and advertisements. The majority of the shows contains prepared speech like news and few conversational speech like interviews. Only 15% of the corpus is narrow band speech. Those data are split in three corpora:

- The training corpus called $ESTER_{train}$ corresponds to 81h (150 shows) composed of 8547 segments in which 3297 full names are detected.

[1] they are officially denoted Phase I and Phase II

- A development corpus[2] corresponds to 12.5h (26 shows) split into 2294 segments containing 920 full names.
- A test corpus called $ESTER_{test}$ contains 10h (18 shows) split into 1417 segments in which 507 full names are detected. it corresponds to the official ESTER evaluation corpus. This corpus contains two radios that are not present in the training corpus. It was also recorded 15 months after the previous data.

This system was competitive: it reached the second position in the ESTER evaluation campaign with 23.6% word error rate (Deléglise et al., 2005; Galliano et al., 2005).

3.1. Acoustic and language models

The vocabulary used by the LIUM system (Deléglise et al., 2005) contained about 65K words. Acoustic models were trained using Ω containing 81h of data with manual transcriptions from four different radios. Those broadcast news are generally wide band but are also composed of phone speech (narrow band). Trigram and quadrigram language models were trained using manual transcriptions of 81 hours of radiophonic broadcast news provided by the ESTER organization resulting in 1.35M words. We add articles from french newspaper "Le Monde" resulting in 319M words.

3.2. Training parameters for confidence measures

Confidence measures are estimated from 4h of same radio stations composing the training corpus for acoustic and language models. We will note this new corpus CTrain. It is independent of the training corpus and we have its manual transcription. We have also an automatic transcription thanks to the recognizer. From the two transcriptions at the same time, we can compute the various features of our measures. For the acoustic confidence measure, we obtain the parameters μ, σ and a of the equation 2. For the LMBB confidence measure, we compute the confidence scores from the error rate obtained with the various LMBB classes (see figure 1).

4. Merging confidence measures

Combining multiple features to result in a single metric can be made in many ways (Schaaf and Kemp, 1997). The most popular used operators are: minimum, maximum, (arithmetic) average, product (or geometric average) and quadratic average. The resulting measure is:

$$m(w) = \mathcal{O}\left(m_1(w), \ldots, m_K(w)\right)$$

where $\mathcal{O}$ is the respective operation. As we have noticed before, our combination rule should respect several constraints. Particularly, the final confidence measure should not alter the global prediction of the average probability of an hypothesis word is correct. The minimum, maximum and product operators do not respect this constraint but can be used when the resulting bias is acceptable.

To take into account the quality of each measure, we can use the weighted average:

$$m(w) = \frac{1}{K}\sum_{i=1}^{N} q_k m_k(w), \text{ with } \sum_{k=1}^{K} q_k = 1.$$

In this case, the weights q_k can be learnt empirically by cross-validation with the results of the obtained measure on the Normalized Cross Entropy (NCE) on CTrain. This metric is used by NIST to assess the quality of a confidence measure and to score evaluations. This is an estimation of how much additional information the confidence tags provide (Evermann and Woodland, 2000; Maison and Gopinath, 2001). Thanks to this metric, we will be able to know the most relevant measures to merge with to obtain a better one.

We can also choose approaches of merging coming from the evidence theory the and probability theory. In fact, we tried this different approaches during our experiments: the approach giving the best results on CTrain in terms of NCE during our experiments on merging the LMMB measure and the acoustic measure probability was a simple linear interpolation: $m_{AC/LMBB}(w) = \nu * m_{AC}(w) + (1-\nu) * m_{LMBB}(w)$. With CTrain, we obtain $\nu = 0.7$.

Table 1 shows that, on CTrain, acoustic and LMBB measures give real information on the word correctness. By merging them, we hope to improve qualities of each one of them. We take the fusion which gives the best results with NCE. This fusion improves the performances of m_{ac} and m_{lmbb}.

Measure	NCE
acoustic $m_{AC}(w)$	**0.035**
LMBB $m_{LMBB}(w)$	**0.072**
fusion $m_{AC/LMBB}(w)$	**0.087**

Table 1: Comparison of confidence measures on CTrain

5. Validation of confidence measures

Before filtering words and thus carrying out a new step in training the acoustic models, we check the influence of the confidence measure on the error rate[3] (deletions are not counted) on the test corpus $ESTER_{test}$.

5.1. Evaluation with reject rate *vs.* error rate

To evaluate the relevance of our fusion measure m (table 1), we put a threshold on the scores to accept only the words which confidence score is higher than it. Thus, we can observe the misrecognized words which should have been rejected among the words we have nevertheless accepted.

Figure 2 shows the rejection and error rates on the test data. The words are accepted/rejected using a threshold on their confidence score with m measure. For example, rejecting

[2] it is the official ESTER phase I development corpus merged with the official ESTER phase II development corpus

[3] in this paper, 'errors rate' term refers to insertions and substitutions, whereas 'word error rate' refers to the common metric refering to insertions, substitutions and deletions. We need to make this distinction to study only recognized words

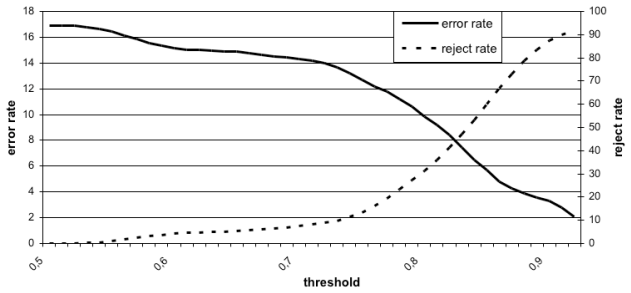

Figure 2: Rejection rate and error rate for the accepted words on $ESTER_{test}$ according to the threshold value.

33% of the words with a threshold at 0.82 means an improvement of the WER from approximately 42% in relative.

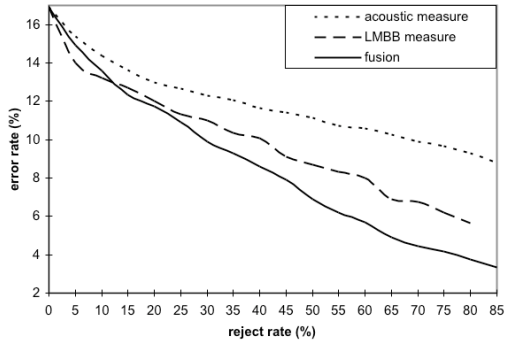

Figure 3: Error rate *vs.* Reject rate: comparison of confidence measures on $ESTER_{test}$.

Figure 3 shows the comparison between the fusion measure and the two measures which compose it. The fusion measure is more effective: by rejecting less words or the same quantity, this measure reaches a better error rate than each of two measures taken one by one.

6. Word-segment based filtering

In this part, we select the best audio segments recognized from a data set of audio files by applying a threshold to each word confidence score (Wessel and Ney, 2005). We noticed during our experiments that the hypotheses with less error rate were made of high confidence word groups. Moreover, it seems better for training algorithms of the acoustic models to preserve long duration segments. This is why we keep only segments of more than 5 seconds and which words have confidence scores higher than a threshold noted α. Plus, if only one word comes to disturb this high-score group with a score slightly lower than the threshold (less than 4%), we preserve it in the segment.

Indeed, we checked that the acceptance of only one word among a high-score group enables us to preserve more relevant segments with only a slight increase of error rate. For example on test data, table 2 shows that segments of more than 5 seconds and for which all the words have confidence scores higher than 0.7 represents 62.3% of the recognized words. This set of segments is noted SEG1. The

Words	**Acceptation rate**		**Error rate**	
no filtering	*100.0%*		*16.9%*	
with filtering	$\alpha > 0.7$	$\alpha > 0.77$	$\alpha > 0.7$	$\alpha > 0.77$
no Seg	92.0%	79.0%	14.3%	11.8%
SEG1	62.3%	35.7%	11.9%	6.9%
SEG1*	71.4%	42.0%	12.5%	8.8%

Table 2: Acceptation and error rates of recognized words according to the filtering method on test data. SEG1 is the set of words which are included in segments during more than 5 seconds and composed only by words whose the confidence scores are greater than a threshold α. SEG1* has the same constraints as SEG1, but accepts only one word by segment with a score slightly lower than α.

error rate for the words of SEG1 is 11.9% (the entire recognized words have an error rate of 16.9%). Adding segments whose only one word has a confidence score smaller than 0.7, and verifying that this word has a confidence score higher than 0.672 (4% of 0.7), the acceptation rate represent 71.4% of the recognized words, for an error rate of 12.5%. This set of segments is noted SEG1*. We can notice that the constraint on the minimal duration of confident word-segment is restrictive: table 2 shows that according to the same value of α, the acceptation rate of SEG1* is very smaller than the acception rate of SEG1. On the other side, the error rate of recognized words decreased too. This confirms that the confidence score of the context of a word is helpful to process this word.

7. Injecting recognized words into training data

In this part, we select the best audio segments recognized on a new data set of 54h from various radio stations by applying a the confident word-segment-based filtering described above. These segments are added to the training corpus of acoustic models, getting then a more significant training corpus for parameters estimation.

The method to choose the value of threshold α_x for filtering is simple: α_x is the value which allows to obtain an acceptation rate of $x\%$. We have used this way to avoid to choose a value too close to the development corpus used to tune parameters of confident measures. So we expect that the more we reject recognized words, the more the error rate of accepted words is low, as shown in results in sections 5.1. and 6.. Here, we cannot evaluate the error rate of accepted words because manual transcriptions are not avalaible for this 54h of data.

Table 3 shows the impact of injecting segments on training corpus of acoustic models in terms of word error rate: the LIUM speech recognition system was used on the ESTER test data using the different acoustic models trained. It is shown in table 3 that injecting 27 hours of non-filtered data into the train corpus of the acoustic model doesn't decrease the word error rate comparing to the acoustic model trained only with the initial corpus. Injecting all the recognized segments (54h of data) allows only a gain of 0.1% absolute (0.42% relative) which is not an improvement statisticaly

significant according to the 95% confidence interval: computed with a word error rate of 23.7% and a set of data composed by 114,000 words, the 95% confidence interval is included in [23.45; 23.95]. Results with a word error rate included in this interval are considered as non-statistically significant ones (Simonin et al., 1998).

Training corpus	**WER**
Initial 81h ($ESTER_{train}$)	23.7%
$ESTER_{train}$ + 27h non-filtered	23.7%
$ESTER_{train}$ + 54h non-filtered	23.6%
$ESTER_{train}$ **+ 28h filtered (α_{50})**	**23.3%**
$ESTER_{train}$ **+ 11h filtered (α_{20})**	**23.4%**

Table 3: Final WER on test data according to injected segments on training corpus of acoustic models

Injecting about the best 50% automatically transcribed data in terms of confident measure allows a gain of 0.4% absolute (1.69% relative) word error rate which is a statisticaly significant result. More interesting, injecting only the best 20% automatically transcribed data improve the word error rate too. These results show the importance of the quality of the injected data compared to their amount. Another interesting result is the fact that the filtering concerns narrow- and wide- band both as shown in table 4.

Training corpus	**Narrow band**	**Wide band**
Initial 81h (Ω)	12h	69h
Ω **+ 28h filtered (α_{50})**	15h (+25%)	94h (+36%)
Ω **+ 11h filtered (α_{20})**	14h (+17%)	78h (+13%)

Table 4: Repartition of training corpus according to the bandwidth

8. Improvements

After a preliminary study for automatic detection of well recognized words with confidence measures which are easilly computable and not affected by the search space size, we try to improve the results by merging measures with word posteriors. As they come from different parts of the system, they should offer complementary information about the word correctness.

8.1. Word posterior probability

Word posterior probabilities can be computed from N-best lists (Stolcke et al., 1997), word-lattices (Evermann and Woodland, 2000) or confusion networks (Mangu et al., 2000). Roughly, the word posterior probability is the ratio of the *a priori* probability of a word and the sum of the *a priori* probabilities of all the alternatives. These *a priori* probabilities are given by a combination of values given by acoustic and language models. Thus, word posteriors can be seen as a summarization of acoustic scores, linguistic scores and search space topology.

In N-best lists, the word posterior probability of a word is approximated with the ratio of the sum of the *a priori* probabilities of the occurrences of this word in the N hypotheses in a given position, and the sum of all the *a priori* probabilities of occurrences of words in this same position, including occurrences of the given word.

In word-lattices- and confusion networks- based approaches, the word posterior probability can be seen as a generalization of the N-best approach, where word-segmentations and search space deepth are better considered.

Unfortunately, this measure is affected by pruning heuristics reducing the size of pruned word-lattices generated during the recognition process and in practice the use of this measure can be biased. To overcome this problem, a decision tree can be trained to transform the posterior probabilities in better confidence scores (Evermann and Woodland, 2000).

In this paper, we use a confusion networks based approach directly derived from (Mangu et al., 2000) to compute word posteriors.

8.2. Fusion

Techniques for merging confidence measures are the same than in section 4. The approach giving the best results terms of NCE on CTrain was also a simple linear interpolation between only the LMBB measure and word posteriors (WP), the acoustic measure is not discriminative enough for word correctness (see table). This fusion measure is called WP/LMBB: $m_{WP/LMBB}(w) = \lambda * m_{WP}(w) + (1-\lambda) * m_{LMBB}(w)$. On CTrain, we obtain, $\lambda = 0.7$.

Measure	**CTrain**	$ESTER_{test}$
acoustic $m_{AC}(w)$	0.022	0.019
LMBB $m_{LMBB}(w)$	0.081	0.063
word posteriors $m_{WP}(w)$	**0.169**	**0.187**
fusion $m_{AC/LMBB}(w)$	0.087	0.072
fusion $m_{WP/LMBB}(w)$	**0.276**	**0.270**

Table 5: Comparison of related confidence measures on train and test data for confidence measure using normalized cross entropy (NCE)

Table 5 shows that on both corpora, the word posterior probability gives real information on the word correctness in a really better way than the single LMBB measure. But, merging this two measures into the WP/LMBB measure provide a very good measure which outperforms the single word posterior probability with a NCE value of 0.270 compared to 0.187 on $ESTER_{test}$. This can be explained by the fact that LMBB and word posterior probability offer complementary information. Plus, we observed a high amount of words for word posteriors scores beyond 0.9 which were not correct. One explanation is that those words do not have much competition in the search space and nevertheless obtain a high word posteriors score. But those word can have a low confidence level according to the LMBB. Thus, the LMBB measure can help word posteriors to be more discriminative.

Figure 4 shows the quality of the WP/LMBB measure. It have the same behaviour of word posteriors but it can reject more words than word posteriors with lower error rate than

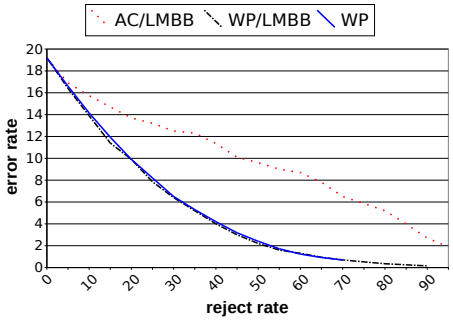

Figure 4: Error rate *vs.* Reject rate: comparison of confidence measures on test data.

the measure AC/LMBB. This new measure let us believe a bigger impact on large scale filtering data.

9. Conclusion

In this article, we propose new confidence measures based on the back-off behavior of language model, a normalization of an existing acoustic confidence measure: their fusion by linear interpolation improves their capacities to detect and reject incorrect words. Separately, thes two first measures are not expensive in computing time, as well as their fusion. We introduce a word-segment -based filtering using the fusion of these confident measures and show that this filtering can extract recognized words with very low rate from automatically transcribed segments in a unsupervised way. In this paper, we use filtered words from audio data with no manual transcription available to increase the size of the training corpus of acoustic models. But other applications exploiting this kind of filtering can be proposed, particularly applications processing the ouput of a automatic speech recognition system (dialog system, named entities extraction, topic detection, ...). Experimental results show a statistically significant improvement of the final word error rate compared to results obtained by using only the initial corpus or by adding data without filtering. A confidence measure improving word posteriors abilities is also presented, improving results of the first fusion. It seems relevant to continue our work to measure the impact of our method on a large scale with our new measure merging word posteriors with an language model-based measure: it should be interesting to know what amount of data filtered in a very restrictive way is necessary to reach the limit of possible improvements.

10. References

L. Chen, L. Lamel, and J.-L. Gauvain. 2004. Lightly supervised acoustic model training using consensus network. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, May.

S. Cox and S. Dasmahapatra. 2002. High-level approaches to confidence estimation in speech recognition. *IEEE Transactions on Speech and Audio Processing*, 10(7).

P. Deléglise, Y. Estève, S. Meignier, and T. Merlin. 2005. The LIUM speech transcription system: a CMU Sphinx III-based system for french broadcast news. In *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, September.

G. Evermann and P.C. Woodland. 2000. Posterior probability decoding, confidence estimation and system combination. In *Speech Transcription Workshop*.

S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.F. Bonastre, and G. Gravier. 2005. The ESTER phase II evaluation campaign for the rich transcription of french broadcast news. In *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, September.

B. Maison and R. Gopinath. 2001. Robust confidence annotation and rejection for continuous speech recognition. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, May.

H. Mangu, E. Brill, and Stolcke A. 2000. Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. *Computer Speech and Language*, pages 4373–400.

F. Metze, T. Kemp, T. Schaaf, T. Schultz, and H. Soltau. 2000. Confidence measure based language identification. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, June.

R. San-Segundo, B. Pellom, K. Hacioglu, W. Ward, and J. Pardo. 2001. Confidence measures for spoken dialogue systems. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, May.

T. Schaaf and T. Kemp. 1997. Confidence measures for spontaneous speech recognition. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, pages 875–878, Munich, Allemagne, April.

J. Simonin, L. Delphin-Poulat, and G. Damnati. 1998. Gaussian Density Tree Structure in a Multi-Gaussian HMM-Based Speech Recognition System. In *Proc. of ISCLP, International Conference on Spoken Language Processing*.

A. Stolcke, Y. Konig, and M. Weintraub. 1997. Explicit word error minimization in N-best list rescoring. In *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, volume 1, pages 163–166, Rhodes, Greece.

C. Uhrick and W. Ward. 1997. Confidence metrics based on n-gram language model backoff behaviors. In *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Rhodes, Greece, September.

F. Wessel and H. Ney. 2005. Unsupervised training of acoustic models for large vocabulary continous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13:23–31.