Confidence Measure Normalization for Robust Selection of ASR Agents

Laurent Couvreur[†], *Jean-Marc Boite*[‡], *Stéphane Dupont*[‡], *Christophe Ris*[‡]

[†]Signal Processing Department, Faculté Polytechnique de Mons, Mons, Belgium
[‡]Speech Recognition Group, Multitel Research Center, Mons, Belgium
couvreur@tctc.fpms.ac.be {boite,dupont,ris}@multitel.be

Abstract

In this paper, we present several normalized confidence measures in hybrid HMM-MLP automatic speech recognition (ASR) systems. We consider purely acoustic confidence measures, *i.e.* computed from frame-level acoustic scores. Such confidence measures are known to be highly sensitive to the training database (*e.g.*, language or lexicon) and the design of the acoustic model (*e.g.*, nature of output states). Hence, it makes tedious setup procedures (*e.g.*, rejection threshold tuning) and does not allows comparing heterogenous ASR systems that are trained on specific databases and operated with proper grammars.

We investigate various normalization approaches such that the resulting acoustic confidence scores are less application dependent. The normalized confidence measures have been applied successfully for run-time selection of the best system among several systems competing together in a collaborative task, namely recognition of command words, natural numbers or letters, in the framework of the IC&C application.

1. Introduction

For the past decades, human-computer interfaces (HCI) have been merely relying on the simple interaction (*e.g.*, keyboard strokes, mouse clicks, screen touches). Nowadays, there is a large effort to integrate integrate new interaction modalities in order to yield interfaces more similar to human-human communication and make their use more intuitive, natural and efficient [1].

Automatic Speech Recognition (ASR) happens to be a promising technology in this framework. As an example, let us mention the IC&C project [2]. Everyone has already seen someone trying to explain his bright new idea by commenting a sketch on a napkin while sitting at a table in a restaurant. The person efficiently combines draw and speech to transmit its message to his interlocutor. The IC&C application consists in a human-computer interface inspired by this communication process. Although the project aims at developing a generic interface as a top layer on any Computer-Aided Design (CAD) software, it has been only applied successfully for architectural design so far (see Figure 1). Unlike classical CAD interfaces that relies on embedded selection menus and complex dialogue boxes, the proposed interface interprets as naturally as possible the user inputs and does not inhibit the design process by setting highly constrained interactions.

The IC&C interface relies on a multi-agent architecture. Every agent is task-oriented and designed to recognize specific graphical or audio data. The user inputs are continuously monitored and agents throw messages when they recognize some data. More especially, agents are grouped into squads. The speech squad consists of several ASR engines that have been trained independently on different speech databases and operate simultaneously on different tasks. For example, there are ASR agents for recognizing specific words to command the interface, numbers to mark dimensions or letters to lay down spelled captions. Depending on the context of the dialogue between the user and the interface, required ASR agents are triggered on incoming utterances and the best recognition is selected afterwards. This flexible approach allows limiting the computational cost and performs better than a single generalist ASR system.

The key issue in a multi-agent ASR approach is the selection procedure. Besides providing a word sequence, every active ASR agent has to score its confidence into its result, the so-called confidence measure. The best word sequence can be obtained by maximizing the confidence measure across the competing ASR agents. To do so, the confidence measures should be homogeneous, that is, their ranges have to be normalized in order to avoid scale effects during comparison. In this paper, we address the problem of normalizing acoustic confidence measures in order to make their ranges uniform for different ASR systems, thereby training databases, acoustic model designs or task grammars.

The paper is organized as follows. In next section, we give a brief description of the ASR systems used in this work. Section 3 defines the confidence measure that is used for ASR agent selection and presents normalization techniques. In section 4, we report selection and recognition results. Conclusions are drawn in section 5.

2. ASR System Overview

Our ASR system [3] relies on the typical architecture presented in Figure 2. It consists of four main blocks. First, the audio interface converts the acoustic wave that is measured by a microphone into a digital speech signal. Second, the front-end (FE) chops the speech signal into frames and computes for each frame a set of acoustic coefficients that capture the essential shape of the power spectrum. In this work, the acoustic coefficients are obtained via the ETSI standardized algorithm for distributed speech recognition [4]. Next, the acoustic coefficient

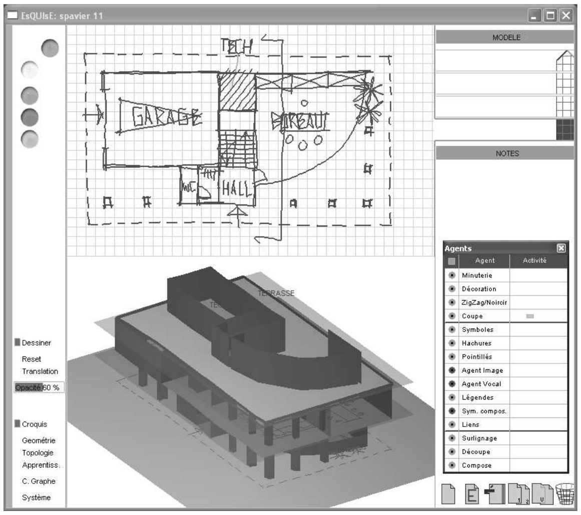

Figure 1: *Screenshot of IC&C application: a natural and user-friendly interface for Computer-Aided Design (CAD) software, for example in architectural design.*

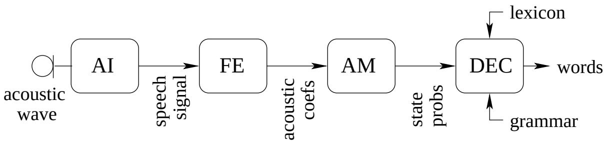

Figure 2: *A typical ASR system: microphone, audio interface (AI), front-end (FE), acoustic model (MA) and word decoder (DEC).*

vectors are fed into the acoustic model (MA). Here, the acoustic model is based on the Multi Layer Perceptron (MLP) / Hidden Markov Models (HMM) paradigm [5]. Such model has to be trained beforehand on large speech databases. The outputs of the MLP estimate the *a posteriori* probabilities of all the HMM states for the acoustic coefficient vectors. Finally, the word decoder (DEC) searches the probability lattice for the most likely word sequence. To do so, every word is represented by a HMM and authorized word sequences are defined in a context-free grammar. The search consists in finding the model sequence, thereby the word sequence, that best fit the acoustic observations within the grammar. It is implemented as a one-pass time-synchronous Viterbi algorithm with pruning [6].

For most ASR tasks, the estimation of a HMM for every word is not practical since it requires a large amount of training material. Hence, the word HMMs are generally obtained by composition of subword HMMs. Most common subword units are context-independent phonemes. With composite word HMMs, it is not required anymore to collect speech databases containing the words to be recognized. The composition is typically based on a lexicon that defines the transcriptions of words into subword HMMs. However, the use of composite word HMMs leads to coarser acoustic modeling. Hence, word-based systems generally outperform subword-based systems, and the former should be preferred when possible.

3. Confidence Measure Normalization

As mentioned in the previous section, the decoder is constrained by a lexicon and a grammar. They actually define all the possible HMM state sequences. For a given utterance, the decoding search consists in finding the best state sequence and the recognized sentence can be viewed as a byproduct of the decoding process.

The confidence in the recognition result can be naturally measured by the accumulated acoustic score along the decoded path [7, 8], that is,

$$\Gamma_1(W) = \Gamma_1(s_1^T) = \frac{1}{T} \sum_{t=1}^{T} \log P(s_t|o_t) \qquad (1)$$

where $s_1^T = \{s_1, \ldots, s_T\}$ is the HMM state sequence estimated by the decoder and corresponding to the recognized sentence W, and $P(s_t|o_t)$ stands for the *a posteriori* probability of being in state s_t given the observed acoustic vector o_t at t-th frame. These probabilities are provided by the MLP and take their value between 0 and 1. Hence, the confidence measure of equation (1) ranges from 0 and $-\infty$ denoting high and low confidence, respectively. Note that the time normalization factor $1/T$ makes the confidence measure to be independent on the utterance length. Otherwise, longer utterances would always lead to lower confidence measures.

Silence intervals are generally very well recognized. Consequently, the confidence measure tends to be abnormally high, *i.e.* too optimistic, when the utterance contains long intervals of silence. Therefore, it is classical to remove silence frames from the computation of the confidence measure. Equation (1) becomes

$$\Gamma_2(W) = \Gamma_2(s_1^T) = \frac{1}{T} \sum_{t=1}^{T} I(s_t) \log P(s_t|o_t) \qquad (2)$$

where the indicator function $I(\cdot)$ is equal to 0 if the state s_t at the t-th frame is the silence state, and 1 otherwise.

The confidence measured based on *a posteriori* state probabilities can be further improved by dividing each probability term in the sum by the best probability for the corresponding time instant, that is,

$$\Gamma_3(W) = \Gamma_3(s_1^T) = \frac{1}{T} \sum_{t=1}^{T} I(s_t) \log \left(\frac{P(s_t|o_t)}{\max\limits_{s_t} P(s_t|o_t)} \right). \qquad (3)$$

It has been shown that this normalization yields very good detection of out-of-vocabulary words [8]. This result holds for detecting out-of-grammar sentences if silence frames are discarded. In the framework of selecting an ASR agent as described in the introduction, we can assume that the right model generally provides a good result. Hence, the selection problem can be viewed as several out-of-grammar sentence detection problems, *i.e.* all the wrong ASR agents should provide low confidence measures to indicate that the utterance is not authorized by their grammar.

In order to reduce further the variability of the confidence measure across ASR systems, we propose to normalize the cu-

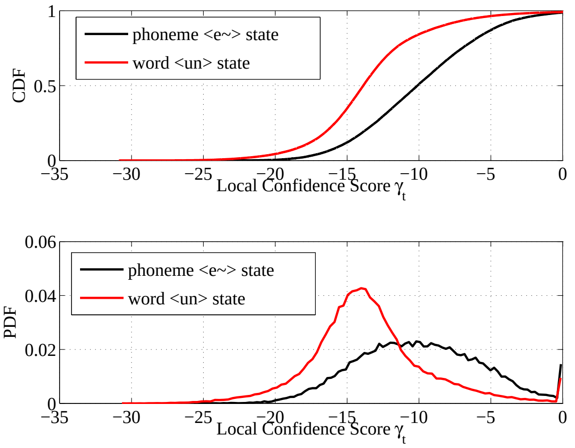

Figure 3: *Cumulative distribution function (CDF) and probability density function (PDF) of local confidence score γ for states of two different acoustic models.*

mulative distribution function of the local confidence scores involved in its computation. Define γ_t as the local confidence score at t-th frame, that is,

$$\gamma_t = \log\left(P(s_t|o_t)/\max_{s_t} P(s_t|o_t)\right). \tag{4}$$

During the training procedure where the state segmentation of the speech training database is known, one can build the cumulative distribution function of the local confidence score for every state of the acoustic model. Define $F_i(\gamma)$ as the cumulative distribution function of the local confidence score for the i-th state. We propose to compute the confidence measure as the average of the cumulative distribution function outputs $F_{s_t}(\gamma_t)$ instead of the local confidence scores γ_t,

$$\Gamma_4(W) = \Gamma_4(s_1^T) = \frac{1}{T}\sum_{t=1}^{T} I(s_t)F_{s_t}(\gamma_t). \tag{5}$$

The cumulative distribution function takes its values between 0 and 1 while its argument γ can have various ranges depending on the state. Hence, the outputs of the cumulative distribution functions can be viewed as normalized local confidence scores that are better candidates for computing a more homogeneous and less sensitive confidence measure. Figures 3 and 4 shows the cumulative distribution function and the probability density function of the local confidence score γ and its normalized version $F(\gamma)$, respectively, for states of two different acoustic models. We clearly see that the functions are significantly different in the former case, while they tend to be similar and uniform in the latter case.

In order to limit the computational cost and the memory footprint of the normalization process, the cumulative distribution functions are represented as parametric functions. More especially, each function is approximated by a sigmoid,

$$\hat{F}_i(\gamma) = \frac{1}{1+e^{-\beta_i(\gamma-\alpha_i)}}. \tag{6}$$

The shift parameter α_i and the steepness parameter β_i of the sigmoid approximating F_i are estimated via a nonlinear least-squares method on the training databases, more exactly a Levenberg-Marquardt optimization algorithm [9] is used,

$$(\alpha_i, \beta_i) = \arg\min_{\alpha,\beta} \sum_{\gamma_t|s_t=i}\left(F_i(\gamma_t) - \frac{1}{1+e^{-\beta_i(\gamma_t-\alpha_i)}}\right)^2. \tag{7}$$

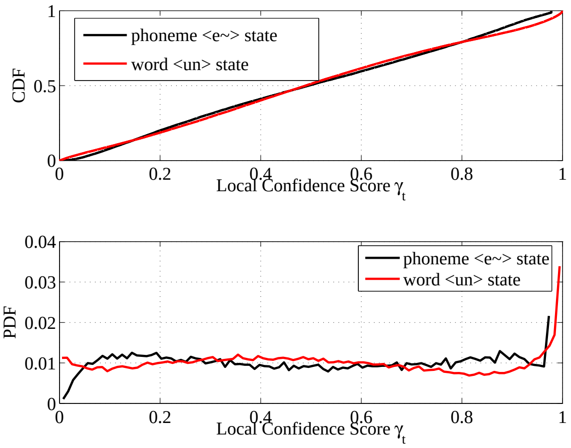

Figure 4: *Cumulative distribution function (CDF) and probability density function (PDF) of normalized local confidence score $F(\gamma)$ for states of two different acoustic models.*

Table 1: *Performance of ASR agents in terms of word error rate (WER) and sentence error rate (SER).*

ASR Agent	Test Set	WER [%]	SER [%]
A	*Commands*	12.0	12.0
B	*Numbers*	1.2	3.5
C	*Letters*	3.7	6.8
B'	*Numbers*	4.5	10.9
C'	*Letters*	25.3	34.6

4. ASR Agent Selection

The experimental framework in this work can be described as follows. We consider three ASR agents, referred as A, B and C in the following. These agents are designed to recognize isolated command words, natural numbers between 0 and 9999, and sequences of letters, respectively. All tasks are in French.

The acoustic model of the command agent consists of phoneme-based HMMs whose state probabilities are estimated by a MLP trained on the BREF80 database [10]. The acoustic model of the number agent and the letter agent consist of word-based HMMs and the corresponding MLPs were trained on parts of the EUROM1 database [11] and the BDSONS database [12].

We first evaluated the performance of the different ASR agents on their specific task (see Table 1). For this purpose, test sets have been collected for every task, namely *commands*, *numbers* and *letters*. Note that the recordings have been made in real conditions and are sometimes corrupted by some background noise. Table 1 also reports the results for two other

Table 2: *Comparison between all-in-one ASR system and selection methods in terms of word error rate (WER) and sentence error rate (SER). Word error rate are detailed as substitution, deletion and insertion error rates.*

Method	Test Set					
	Commands		*Numbers*		*Letters*	
	WER [%]	SER [%]	WER [%]	SER [%]	WER [%]	SER [%]
All-in-one	31.1 (14.6/5.4/11.4)	20.0	5.3 (1.8/3.3/0.1)	12.1	27.8 (18.4/7.6/1.9)	39.9
Γ_1*-based selection*	70.8 (25.9/0.3/44.5)	26.3	5.7 (3.0/1.6/1.0)	11.0	18.4 (15.9/1.8/0.6)	36.4
Γ_2*-based selection*	35.6 (21.3/0.1/14.3)	21.4	5.3 (3.3/1.0/1.0)	8.7	19.5 (8.3/10.7/0.5)	19.7
Γ_3*-based selection*	26.4 (18.8/0.1/7.6)	18.9	2.8 (1.2/0.8/0.8)	6.1	12.4 (6.4/5.5/0.6)	16.0
Γ_4*-based selection*	20.4 (16.4/0.2/3.8)	19.0	1.2 (0.1/0.4/0.7)	3.5	8.7 (4.2/3.7/0.8)	11.2
Oracle selection	12.0 (11.8/0.2/0.0)	12.0	1.2 (0.1/0.4/0.7)	3.5	3.7 (1.6/1.3/0.8)	6.80

agents B' and C'. These agents use phoneme-based HMMs with the MLP of agent A but the same grammar as agents B and C, respectively. As mentioned in section 2, we see that the word-based agents perform significantly better than their equivalent phoneme-based agents. This can be explained by the better modeling of in-word parts of speech in the former case.

Table 2 reports recognition results when selection is applied at run-time. This selection consists is running the ASR agents all together and keep the result provided by the one with the highest confidence measure. Various confidence measures as defined by equations (1) to (5) were tested. We first observe that the selection approaches always perform better than the *all-in-one* method that consists of a generalist acoustic model and a global grammar encompassing all the agent grammars. This observation justifies itself the use of a squad of specialized ASR agents instead of a generic model. Next, we observe that the selection based on the confidence measure Γ_4 yields the best performance. Besides, it gets close and even equal for the *numbers* test set to the *oracle* selection, which consists in always selecting the right ASR agents.

5. Conclusions

In this work, we presented several approaches to reduce the variability of acoustic confidence measures based on state *a posteriori* probabilities. More especially, we proposed a method that normalizes the statistical distribution of the confidence measure values across acoustic models and grammars. The normalized confidence measure has been successfully applied for selecting the best system among several specialized ASR agents competing together in a common recognition task. We showed that the selection-based approach significantly outperforms the *all-in-one* system, which consists of a generalist ASR system.

6. Acknowledgments

This work has been partly supported by the Direction Générale des Technologies, de la Recherche et de l'Energie (DGTRE) of the gion Wallonne (RW) in Belgium, under contract number WIST-03/1/5593 (project IC&C).

7. References

[1] SimilarNet, "The European Taskforce for Creating Human-Machine Interfaces Similar to Human-Human Communication", `http://www.similar.cc`.

[2] IC&C, "A Creative Interface for Design", `http://tcts.fpms.ac.be/research.php`.

[3] J.-M. Boite, L. Couvreur, S. Dupont and C. Ris, "Speech Training and Recognition Unified Tool (STRUT)", `http://tcts.fpms.ac.be/asr`.

[4] ETSI ES 202050, "Speech Processing, Transmission and Quality Aspects (STQ), Distributed Speech Recognition, Advanced Front-End Feature Extraction Algorithm, Compression Algorithm", *ETSI Standard*, Jul. 2002.

[5] H. Bourlard and N. Morgan, "Connectionist Speech Recognition – A Hybrid Approach", *Kluwer Academic Publisher*, 1994.

[6] X. Huang, A. Acero and H.-W. Hon, "Spoken Language Processing: A Guide to Theory, Algorithm, and System Development", *Prentice Hall*, pp. 522–525, 2001.

[7] G. Williams and S. Renals, "Confidence Measures from Local Posterior Probability Estimates", *Computer Speech and Language*, vol. 13, no. 4, pp. 395–411, Oct. 1999.

[8] E. Mengusoglu and C. Ris, "Use of Acoustic Prior Information for Confidence Measure in ASR Applications", *Proc. EUROSPEECH*, pp. 2557–2560, Aalbord, Denmark, Sep. 2001.

[9] J. E. Dennis, "Nonlinear Least-Squares", ed. D. Jacobs, "State of the Art in Numerical Analysis", *Academic Press*, pp. 269–312, 1977.

[10] L. F. Lamel, J. L. Gauvain and M. Eskénazi, "BREF, a Large Vocabulary Spoken Corpus for French", *Proc. EUROSPEECH*, pp. 505–508, Geneva, Italy, Sep. 1991.

[11] J. Zeiliger, J.-F. Serignat, D. Autesserre and J.-M. Dolmazon, "EUROM1: une Base de Données Parole Multilingue", *Proc. JEP*, pp. 303–306, Bruxelles, Belgique, May. 1992.

[12] R. Carré, R. Descout, M. Eskénazi, J. Mariani and M. Rossi, "The French Language Database: Defining, Planning and Recording a Large Database", *Proc. ICASSP*, pp. 324–327, San Diego, California, Mar. 1984.