A DISCRIMINATIVE TRAINING FRAMEWORK USING N-BEST SPEECH RECOGNITION TRANSCRIPTIONS AND SCORES FOR SPOKEN UTTERANCE CLASSIFICATION

[1]*Sibel Yaman,* [2]*Li Deng,* [2]*Dong Yu,* [2]*Ye-Yi Wang,* [2]*Alex Acero*

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA
[2]Microsoft Research, Microsoft Corporation, Redmond, WA
[1]syaman@ece.gatech.edu
[2]{deng, dongyu, yeyiwang, alexac}@microsoft.com

ABSTRACT

In this paper, we propose a novel discriminative training approach to spoken utterance classification (SUC). The ultimate objective of the SUC task, originally developed to map a spoken speech utterance into the most appropriate semantic class, is to minimize the classification error rate (CER). Conventionally, a two-phase approach is adapted, in which the first phase is the ASR transcription phase, and the second phase is the semantic classification phase. In the proposed framework, the classification error rate is approximated as differentiable functions of the language and classifier model parameters. Furthermore, in order to exploit all the available information from the first phase, class-specific discriminant functions are defined based on score functions derived from the N-best lists. Our experimental results on the standard ATIS database indicate a notable reduction in CER from the earlier best result on the identical task. The proposed framework achieved a reduction of CER from 4.92% to 4.04%.

Index Terms— spoken utterance classification, discriminative training, automatic speech recognition, statistical language modeling.

1. INTRODUCTION

Due to the tremendous progress in automatic speech recognition (ASR) technology in the past decades, extensive research has been devoted to its potential commercial applications. Among these, spoken utterance classification (SUC), a special form of spoken language understanding, has found many practical applications including call routing [1], dialog systems and command and control [2]. Whether a speaker is inquiring about flights to a city or reserving a table at a restaurant, the ultimate objective of SUC systems is to reduce the classification error rate (CER).

Conventionally, a two-phase approach is adapted for SUC task, in which the first phase is the ASR transcription phase, and the second phase is the semantic classification phase. Typically, an in-domain language model (LM) is trained so that the word error rate (WER) in the ASR phase is reduced. Once the spoken utterance is automatically transcribed, the semantic classification essentially becomes a text classification problem. Therefore, reducing errors in the ASR transcription can improve CER by virtue of providing better transcriptions to the classifier. However, it has been reported that reductions in WER do not necessarily translate into reductions in CER [3, 4]. More important than WER reduction, the models used in SUC task should be trained to consistently match the ultimate objective of CER reduction.

In this paper, we describe a novel discriminative training framework for learning the models used for SUC task, specifically for training the language and classifier models. Based on the minimum classification error rate minimization (MCE) principles [5], the CER is approximated as differentiable functions of the LM and classifier parameters. The major contribution of this paper is the way the N-best hypotheses generated by ASR module are used in discriminative training of the model parameters. More precisely, the proposed framework associates a score to each pair formed with the sentences in the N-best list and the semantic classes. This score represents how likely it is that a sentence yields classification of the spoken utterance into its paired category. These scores also behave as the class-specific discriminant functions in discriminative training. We illustrate the use of the proposed framework on ATIS database. We observed a significant improvement over the earlier best system on the identical task. Our experiments indicate a CER reduction from 4.92% to 4.04%. These findings suggest that when WER is below certain level, the CER attained by using ASR transcriptions can be lower than the CER attained by using manual transcriptions.

This paper is organized as follows: In Section 2, the proposed framework for discriminative training of the language and classifier models using the N-best lists is described. In Section 3, some practical methods for successful implementation of the proposed framework are discussed. Our experimental results using the ATIS set are reported in Section 4. Finally, the concluding remarks and future research directions are investigated in Section 5.

2. A NEW DISCRIMINATIVE TRAINING FRAMEWORK FOR SPOKEN UTTERANCE CLASSIFICATION

Given a spoken speech utterance X_r and a set of M semantic classes $\mathcal{C} = \{C_1, ..., C_M\}$, a spoken utterance classification (SUC) system maps X_r into a class $\hat{C}_r$ so that:

$$\hat{C}_r = \arg\max_{C_r} P(C_r|X_r). \tag{1}$$

Most practical systems make the assumption that the solution to this classification task is a two-phase process. First, a speech recognizer converts X_r into the best-matching sentence W_r. A classifier, then, maps W_r to a semantic class $\hat{C} \in \mathcal{C}$. In most of the conventional SUC approaches, the acoustic and language models of the first phase are trained so that the WER in the ASR phase is reduced. Reducing errors in the automatic transcription can improve CER by virtue of providing better transcriptions to the semantic classifier. However, it has been reported that reductions in WER do not necessarily translate into reductions in CER. In many cases, a user's utterance contains multiple salient phrases pointing a single interpretation of the utterance. Therefore, as long as enough words are recognized to trigger the correct salient phrase, the correct meaning is assigned to the utterance.

In [3], Wang et al. investigate how good the WER is as an indicator of the CER in SUC systems. Their prominent conclusion is that the model training criteria that match the optimization objective are *at least* as important as reductions in WER. For this reason, in our framework, the resulting corpus of transcribed word strings is used not only to train the classifier model but also to train application-specific LMs to be used by the ASR module. Hence, our proposed framework offers better *recognition for semantic classification* models that optimize the ultimate goal of reducing CER in SUC systems.

2.1. DT Framework Using the N-Best Lists

Suppose that W_r is a sentence with probability $P(W_r)$, and L is the LM weight. Further, let $P(C_r|W_r)$ denote the probability of mapping W_r into semantic class C_r, and $P(X_r|W_r)$ denote the probability of transcribing X_r as W_r. Then, the classification decision rule given in Eq. (1) can be approximated as follows:

$$\begin{aligned} \hat{C}_r &= \arg\max_{C_r} \left[\sum_{W_r} P(C_r|W_r, X_r) P(X_r|W_r) P(W_r) \right] \\ &\cong \arg\max_{C_r} \max_{W_r \in \aleph} P(C_r|W_r) P^{\frac{1}{L}}(X_r|W_r) P(W_r). \end{aligned} \quad (2)$$

The approximation in the second line is such that the sentences W_r are limited to the ones in the N-best list, and the term with maximum contribution replaces the summation over all possible W_r. Assume that the semantic classifiers are modeled using the maximum entropy principle, which yields:

$$P(C_r|W_r) = \frac{1}{Z(W_r)} \exp \left[\sum_i \lambda_i f_i(C_r, W_r), \right] \quad (3)$$

where λ_is denote the classifier parameters, f_is denote lexical n-gram feature functions, and $Z(W_r)$ is a normalization term given by:

$$Z(W_r) = \sum_{C_r} \exp \quad \sum_i \lambda_i f_i(C_r, W_r). \Bigg) \quad (4)$$

Examples of f_i are binary bigram features in the following form:

$$f^{bigram}_{c,w_x w_y}(C_r, W_r) = \begin{matrix} 1, \text{ if } C_r = c \wedge w_x w_y \in W_r, \\ 0, \text{ otherwise.} \end{matrix}$$

$P(X_r|W_r)$ is obtained from the speech lattice by summing up the acoustic scores of all the paths that yield W_r for X_r. Let $D(C_r, W_r; X_r)$ be defined as:

$$D(C_r, W_r; X_r) = \log[P(C_r|W_r) P^{\frac{1}{L}}(X_r|W_r) P(W_r)]. \quad (5)$$

Then, Eq. (2) can be replaced with its equivalent (logarithmic) form as:

$$\hat{C}_r \cong \arg\max_{C_r} \max_{W_r \in \aleph} D(C_r, W_r; X_r) \ . \quad (6)$$

$D(C_r, W_r; X_r)$ is called the class-discriminant function. $D(C_r, W_r; X_r)$ shows how likely it is that the sentence W_r yields the semantic class C_r for given X_r. Therefore, for given X_r, $D(C_r, W_r; X_r)$ serves as a *joint association score* between C_r and W_r.

In the training stage, the sentence W_r^0 that is most likely to yield the correct semantic class C_r^0 is *first* extracted based on the $D(.)$ scores[1]. This corresponds to:

$$W_r^0 = \arg\max_{W_r \in \aleph} D(C_r^0, W_r; X_r) \ . \quad (7)$$

Note that the sentence W_r^0 is extracted among all the sentences in the N-best list. Hence, it is the most likely sentence to yield the correct decision for X_r *independent of* whether or not it has the most correct transcription. Furthermore, it is possible to assign the remaining sentences in the N-best list to different semantic classes in a similar manner. For $n = 1, ...,$ let $\mathcal{C}^n$ denote the set of semantic classes and $\aleph^n$ denote the set of the sentences that are not yet assigned. $\mathcal{C}^n$ and $\aleph^n$ are more formally defined as $\mathcal{C}^n = \mathcal{C} \setminus \{C_r^1, ..., C_r^{n-1}\}$ and $\aleph^n = \aleph \setminus \{W_r^1, ..., W_r^{n-1}\}$, where $\setminus$ denotes set-difference. The classes in $\mathcal{C}^n$ and the sentences in $\aleph^n$ are paired with each other by:

$$C_r^n = \arg\max_{C_r \in \mathcal{C}^n} \max_{W_r \in \aleph^n} D(C_r^n, W_r; X_r) \ . \quad (8)$$

For each C_r^n, the corresponding sentence W_r^n is the term inside the square-brackets, i.e.,

$$W_r^n = \max_{W_r \in \aleph^n} D(C_r^n, W_r; X_r). \quad (9)$$

This results in $T = \min\{M, N\}$ such (W_r^j, C_r^j) pairs. Apparently, this kind of assignment of sentences in the N-best list to different classes in $\mathcal{C}$ is an effective mechanism for discriminating the sentence in the N-best list that is most likely to yield the correct class from those that are *more likely* to yield other (wrong) classes.

Upon defining the discriminant functions, a class-specific misclassification function $d_r(X_r)$ and loss function $\ell_r(d_r(X_r))$ are assigned to each X_r, where

$$d_r(X_r) = -D(C_r^0, W_r^0; X_r) + \log \left[\frac{1}{T} \sum_{n=1}^{T} \exp[\eta D(C_r^n, W_r^n; X_r)] \right]^{\frac{1}{\eta}}, \quad (10)$$

$$\ell_r(d_r(X_r)) = \frac{1}{1 + \exp(-\alpha d_r(X_r) + \beta)}. \quad (11)$$

The loss function $\ell_r(d_r(X_r))$ emulates the classification decision loss for each utterance: When $d_r(X_r)$ is very small (or, very large), the loss of classifying it into the class C_r^0 is 0 (or, 1). Given the LM model Λ_W, and the semantic classifier model Λ_λ, the total loss is then approximated as:

$$L(\Lambda_W, \Lambda_\lambda) = \sum_r \ell_r(d_r(X_r)) \quad (12)$$

Being able to approximate the total classification loss as continuously differentiable functions of model parameters make it possible to find the models that minimize the classification loss.

2.2. An Illustrative Example

To clarify our formulation, we next illustrate the assignment of different C_r to different W_r in Table 1. In the training stage, the sentence W_r^0 that yields the highest $D(.)$ with the correct class C_r^0 (here, GRD_SRV) is first extracted. Then, the sentence other than

[1] Throughout our discussion, we reserve the superscript ”0” for the correct class and its associated sentence, whereas the superscript ”n” is reserved for the wrong classes and their associated sentences.

C_r (Class)	W_r (Corresponding sentence)	$D(.)$
GRD_SRV	What is the ground transportation in Atlanta	
C_r^0:GRD_SRV	W_r^0:What is the transportation in Atlanta	-20.04
C_r^1:FARE	W_r^1:What is the round trip fare from Atlanta	-17.56
C_r^2:CITY	W_r^2:What is the transportation Atlanta	-25.46
C_r^3:FLIGHT	W_r^3:What is the transportation and Atlanta	-28.49
C_r^4:FARE_BS	W_r^4:What is the round trip fare from the Atlanta	-27.98
C_r^5:AIR_SRV	W_r^5:What is the transportation the Atlanta	-29.09

Table 1. Assignment of semantic classes to sentences along with the corresponding joint scores.

W_r^0 that has the highest $D(.)$ with any class other than C_r^0 is extracted, which are denoted as C_r^1 and W_r^1, respectively. Next, the sentence other than W_r^0 and W_r^1 that has the highest $D(.)$ with any class other than C_r^0 and C_r^1 is found. This procedure is repeated until either all sentences in the N-best list or all the classes in $\mathcal{C}$ are assigned. In the test stage, X_r is mapped into the class $\hat{C}_r$ that has the highest $D(.)$ with any sentence in the N-best list.

In this example, the joint association score of W_r^0 with C_r^0 is -20.04 . On the other hand, the joint association score of W_r^1 with C_r^1 is -17.56. Hence, if X_r were used in the test stage, it would be assigned to the class `Fare`, which would imply a misclassification for the spoken utterance, X_r.

2.3. Discriminative Training of LM Parameters

Following the general MCE training principles, the LM probabilities that minimize the total loss function $L(\Lambda_W, \Lambda_\lambda)$ can be learned by finding where its gradient vanishes. Doing so results in rules for updating the lexical n-gram (log)probabilities.

For example, let $n(W_r^0, w_x w_y)$ denote the number of times the bigram $w_x w_y$ appears in W_r^0, and $n(W_r^n, w_x w_y)$ denote the number of times the bigram $w_x w_y$ appears in W_r^n. Then, for updating the bigram log probability $p_{w_x w_y} = P(w_y|w_x)$, we eventually get:

$$\begin{aligned} p_{w_x w_y}^{(t+1)} &= p_{w_x w_y}^{(t)} - \varepsilon_{LM} \sum_r \frac{\partial \ell_r(d_r(X_r))}{\partial p_{w_x w_y}} \qquad (13) \\ &= p_{w_x w_y}^{(t)} - \varepsilon_{LM} \alpha \sum_r \ell_r(d_r)[1-\ell_r(d_r)] \frac{\partial d_r(X_r)}{\partial p_{w_x w_y}}, \end{aligned}$$

where

$$\frac{\partial d_r(X_r)}{\partial p_{w_x w_y}} = -n(W_r^n, w_x w_y) + \sum_{n=1}^{N} H_r^n n(W_r^n, w_x w_y). \qquad (14)$$

The weighting coefficients H_r^n are given by:

$$H_r^n = \frac{\exp[\eta D(C_r^n, W_r^n; X_r)]}{\sum_{m=1}^{T} \exp[\eta D(C_r^m, W_r^m; X_r)]}. \qquad (15)$$

With a closer look at Eq. (14), it is noticed that when $\eta \rightarrow \infty$, only the correct (W_r^0, C_r^0) and the most competitive (W_r^1, C_r^1) hypotheses take role in updating $p_{w_x w_y}$. The LM parameters corresponding to the bigrams that are present in W_r^0 but not in W_r^1 (in the example above, *"the transportation"*, *"transportation in"*, *"in Atlanta"*) are increased. In contrast, the LM parameters corresponding to the bigrams found in W_r^1 but not in W_r^0 (*"the round"*, *"round trip"*, *"trip fare"*, *"fare from"*, *"from Atlanta"*) are decreased. The updates for the bigrams common to both W_r^0 and W_r^1 (*"what is"*, *"is the"*) cancel each other, and the corresponding LM parameters are left unchanged. Similar arguments are valid for other lexical n-grams as well.

2.4. Discriminative Training of Classifier Parameters

The semantic classifier model parameters $\lambda_k, k = 1, ..., K$, that minimize the total loss $L(\Lambda_W, \Lambda_\lambda)$ are also learned based on the MCE principles. We get the following update rules for the classifier parameters:

$$\begin{aligned} \lambda_k^{(t+1)} &= \lambda_k^{(t)} - \varepsilon_\lambda \sum_r \frac{\partial \ell_r(d_r(X_r))}{\lambda_k} \qquad (16) \\ &= \lambda_k^{(t)} - \varepsilon_\lambda \alpha \sum_r \ell_r(d_r(X_r))[1-\ell_r(d_r(X_r))] \frac{\partial d_r(X_r)}{\partial \lambda_k}, \end{aligned}$$

where

$$\frac{\partial d_r(X_r)}{\partial \lambda_k} = \varphi(C_r^0, W_r^0) + \sum_{j=1}^{T} H_r^j \varphi(C_r^j, W_r^j). \qquad (17)$$

The term $\varphi_k(C_r^j, W_r^j)$ stands for the derivative of $D(C_r^j, W_r^j; X_r)$ with respect to λ_k, and is given by:

$$\begin{aligned} \varphi_k(C_r^j, W_r^j) &= \frac{\partial \log P(C_r^j|W_r^j)}{\partial \lambda_k} \qquad (18) \\ &= f_k(C_r^j, W_r^j) - \sum_{\tilde{C}} \frac{\exp\left[\sum_i \lambda_i f_i(\tilde{C}, W_r^j)\right]}{\sum_{\hat{C}} \exp\left[\sum_i \lambda_i f_i(\hat{C}, W_r^j)\right]} f_k(\tilde{C}, W_r^j). \end{aligned}$$

Similarly to the case with the update of LM parameters, when $\eta \rightarrow \infty$, the classifier parameters λ_k associated with the bigrams that are present in W_r^0 but not in W_r^1 are increased. In contrast, the classifier parameters λ_k associated with the bigrams that are present in W_r^1 but not in W_r^0 are decreased. Those parameters associated with the same bigrams are left unchanged as before.

3. PRACTICAL ISSUES

3.1. Adjusting the Sigmoid Parameters

One issue that arises is the adjustment of the sigmoid parameters in the loss function $\ell_r(d_r(X_r))$ in Eq. (11). α is a positive constant that controls the size of the learning window and the learning rate, and β is a real number measuring the offset of $d_r(X_r)$ from 0. In our implementation, α is fixed at certain value for all classes, whereas β_j is adjusted class-specifically.

Let ϕ be such that $\phi \geq 0$. Let Ω_j^+ denote the set of all X_r that belong to the class C_j, and Ω_j^- denote the set of all X_r that do not belong to the class C_j. Also, let Υ_j^+ denote the set of all $X_r \in \Omega_j^+$ such that $\Upsilon_j^+ = \{X_r : d_r(X_r) < \phi\mu_j^+\}$, and likewise, Υ_j^- denote the set of all $X_r \in \Omega_j^-$ such that $\Upsilon_j^- = \{X_r : d_r(X_r) > \phi\mu_j^-\}$. Then, β for the j^{th} semantic class is set to:

$$\beta_j = \frac{\beta_j^{pos} + \beta_j^{neg}}{2}, \qquad (19)$$

where β_j^{pos} and β_j^{neg} are the average $d_r(X_r)$ when $X_r \in \Upsilon_j^+$ and $X_r \in \Upsilon_j^-$, respectively. The idea behind the adapted heuristics is to associate more loss for those samples for which $\ell_r(d_r(X_r))$ is close to 0.5. Such samples represent the more confusable examples for the classifier.

3.2. Stopping Criteria

Optimization theory states that the gradient must vanish in an open neighborhood of any (local) optimal solution. Hence, the numerical

	Test WER	Test CER
Manual Transcription	0.00%	4.81%
ASR Transcription	4.82%	4.92%

Table 2. The performance of the baseline system on text inputs and speech inputs for ATIS domain. Both the classifier and the trigram LM for ASR are trained from the in-domain manual transcriptions.

optimization processes formulated in Eqs. (13) and (16) should be stopped when there is no further change (or insignificant change) in the total loss function, $L(\Lambda_W, \Lambda_\lambda)$. A development is used to make the stopping decision in our numerical optimization implementation.

4. EXPERIMENTS

We used the ATIS database to evaluate the proposed framework with the same experimental setup as in [6]. ATIS2 and ATIS3 Category A data are used for training (5798 utterances), ATIS3 1993 and 1994 Category A test set (914 utterances) for testing, and the ATIS3 development set (410 utterances) for tuning system parameters and the stopping criteria as mentioned in Section 3. In our experiments, we used the recognizer that was provided as part of the Microsoft Speech API (SAPI) without adaptations to its acoustic model.

4.1. Performance of the Baseline System

The baseline system design [6] first requires the extraction of the best-matching sentence W_r for each speech utterance X_r. For doing this, a trigram in-domain LM is trained using the same data for the automatic recognition of the test utterances. In the second phase, maximum-entropy classifiers are trained using the manually transcribed training data. The best-scenario ASR word error rates (WERs) and classification error rates (CERs) for the baseline system are tabulated in Table 2. These results were the best on this standard SUC task in the literature prior to the work described in this paper.

4.2. Performance of the Proposed DT Framework using N-best ASR Transcriptions

We now report the experimental evaluation of our proposed framework. A trigram LM based on n-gram frequencies is used as the initial LM. We then train an LM as described in Section 2.3. The classifiers are initialized with classifiers trained with the maximum entropy criterion. This is followed by the classifier training as described in Section 2.4. Note that finding the optimal the LM and classifier parameters require several ”inner” iterations. Once optimal solutions are reached for the LM, the resulting LM is fed-back into the ASR module. This results in new ASR transcriptions and using these new ASR transcriptions, the entire process is repeated, i.e., new classifiers are trained with maximum entropy criterion, new LMs are trained, and finally, the classifier parameters are trained with the training criterion of minimizing the total loss function, $L(\Lambda_W, \Lambda_\lambda)$. This process is repeated for several ”outer” iterations.

In our experiments, we have set $\eta = 1.0, \varepsilon_{LM} = 0.001, \varepsilon_\lambda = 0.03, \alpha = 0.5, L = 1, \phi = 0.1$. These parameters were experimentally tuned using the development data. The performance of the the proposed system is summarized in Table 3. We have listed the WERs and CERs on both the development data and the test data at each outer iteration. It is rather striking that the CER has been reduced from 4.92%, attained by the baseline system using ASR transcriptions, to 4.04%. Furthermore, the CER 4.04% is significantly lower than 4.81%, which was attained by the baseline system with error-free (manually transcribed) speech utterances. Taken together, the finding in Table 3 indicate that when WER is reduced below certain level, the CER attained by ASR transcriptions can be lower than the CER attained by the manual transcriptions.

iteration	Dev. WER	Dev. CER	Test WER	Test CER
1	7.3%	5.61%	6.2%	4.60%
2	7.2%	5.61%	6.4%	4.38%
3	7.4%	5.12%	6.4%	4.38%
4	7.2%	5.12%	6.3%	4.27%
5	7.3%	4.63%	6.4%	4.04%
6	7.4%	5.37%	6.5%	4.04%

Table 3. The proposed DT framework improves CER improves. Significant improvement over the baseline system is achieved.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we described a new discriminative training (DT) framework for spoken utterance classification (SUC). The key novelty is the development of an integrated learning framework for the language model (LM) and the semantic classifier model parameters based on the N-best lists generated in the ASR phase. The parameter learning is based on the minimization of a smooth approximation of the classification error rate (CER). Our experimental results on the standard ATIS SUC task revealed the achievement of a significant improvement in CER from the earlier best system on the identical task.

We have used scores derived from the N-best ASR transcriptions to define the class-specific discriminant functions. The use of N-best transcriptions is motivated by the fact that the same semantic class is often associated with many variants of spoken utterances. In order to learn reliable models, we need sufficient data that capture such variations and their semantic classes. The N-best transcriptions provide one such source of data.

One direction of future research is an extension of the current implementation to the estimation of other system parameters, such as the HMM and pronunciation model parameters. It is also possible to extend this approach illustrated in the context of SUC to more general spoken language understanding tasks, including slot filling using conditional random fields (CRFs).

6. REFERENCES

[1] H.-K. J. Kuo, I. Zitouni, E. Fosler-Lussier, E. Ammicht, and C.-H. Lee, “Discriminative training for call classification and routing,” in *the Proceedings of International Conference on Speech and Language Processing*, Denver, Colorado, 2002.

[2] J. R. Bellegarda, “Semantic inference: A data driven solution for nl interaction,” in *the Proceedings of International Conference on Speech and Language Processing*, Denver, Colorado, 2002.

[3] Y.-Y. Wang, A. Acero, and C. Chelba, “Is word error rate a good indicator for spoken language understanding?,” in *the Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, Virgin Islands, Dec 2003.

[4] G. Riccardi and A.L. Gorin, “Stochastic language models for speech recognition and understanding,” in *the Proceedings of ICSLP*, Sydney, Australia, 1998.

[5] W. Chou, “Discriminant-function-based minimum recognition error rate pattern-recognition approacht to speech recognition,” *Proceedings of IEEE*, vol. 88, no. 8, pp. 1201–1224, August 2000.

[6] Y.-Y. Wang, J. Lee, and A. Acero, “Speech utterance classification model training without manual transcriptions,” in *the Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, 2006.