INTERSPEECH 2009 BRIGHTON

Unsupervised Estimation of the Language Model Scaling Factor

Christopher M. White, *Ariya Rastrow*, *Sanjeev Khudanpur* and *Frederick Jelinek*

Human Language Technology Center of Excellence, and
Center for Language and Speech Processing, Johns Hopkins University
{cmileswhite,ariya,khudanpur,jelinek}@jhu.edu

10.21437/Interspeech.2009-346

Abstract

This paper addresses the adjustment of the language model (LM) scaling factor of an automatic speech recognition (ASR) system for a new domain using only un-transcribed speech. The main idea is to replace the (unavailable) reference transcript with an automatic transcript generated by an independent ASR system, and adjust parameters using this *sloppy reference*. It is shown that despite its fairly high error rate (ca. 35%), choosing the scaling factor to minimize disagreement with the erroneous transcripts is still an effective recipe for model selection. This effectiveness is demonstrated by adjusting an ASR system trained on Broadcast News to transcribe the MIT Lectures corpus. An ASR system for telephone speech produces the sloppy reference, and optimizing towards it yields a nearly optimal LM scaling factor for the MIT Lectures corpus.

Index Terms: LVCSR, language modeling, domain adaptation, semi-supervised learning, unsupervised learning

1. Introduction

Classifier design decisions, particularly those involving selection of a modest number of parameter values, are usually based on optimizing empirical performance on labeled *held out* data. We investigate the use of unlabeled data for making such decisions. Concretely, we investigate the unsupervised selection of the language model scale factor when porting an automatic speech recognition (ASR) system to a new domain. The general idea we investigate is whether and how well labels assigned to the held out data by another (independent) classifier can replace "true" labels; in case of ASR, this is an automatic transcript of the held out speech produced by another ASR system.

The performance of an ASR system is sensitive to deviations from the conditions in which the system was trained, be it the channel, noise, speaker, speaking style or the application domain. Among others, the language model (LM) scaling factor has been shown to be an important domain-dependent [1, 2] parameter that significantly affects word error rate (WER). The conventional method to adjust (tune) the scaling factor to a new domain is to minimize WER via a line search on a labeled development set. Such development data are sometimes difficult or time consuming to acquire, particularly for deployed/live systems experiencing dynamic genre shifts. In such a setting, we investigate if an ASR system can automatically adjust the scaling factor using transcripts produced by a second ASR system.

Less-than-perfect transcripts have been shown to be effective in estimating other ASR system parameters. The so called *quick transcripts* of the Fisher corpus find extensive use in acoustic model training, even though inter-transcriber agreement is significantly less than 100% [3]. Closed captions of Broadcast audio have been successfully used for *lightly supervised* training. Methods such as the *Mechanical Turk* provide useful, even if low quality, labels in other applications in natural language processing [4]. These methods require humans in the loop, while the method we investigate is fully automated.

Fully automatic transcripts have been used for *self training* of acoustic models [5]. Our method is most similar to self training, but uses externally- instead of self-labeled data. A distinction between self training of acoustic models and our setting is that errors in the transcripts impact acoustic training only indirectly, while their impact here is more direct. E.g. instead of each EM iteration increasing the likelihood of the acoustics under the correct phonemic transcript, the likelihood is increased in self training under a "nearby" phonemic transcript, and it is plausible that the parameter updates also increase the likelihood under the correct transcript substantially, even if not maximally. By contrast, when discriminating different LM scaling factors based on their WER on a held out set, *noise* in the measured WER due to the "reference" transcripts being erroneous could lead to a much more detrimental choice.

We begin in Section 2 with a brief theoretical treatment of the use of a *secondary* classifier to produce the "reference" labels for tuning parameters of the *primary* classifier. We view incarnations of the primary classifier with different parameter values as an collection of distinct classifiers, and suggest a method for choosing among them based on their (dis)agreement with the output of the secondary classifier.

We then describe in Section 3 an experiment to test the method by tuning, to a Lecture domain, the LM scaling factor of a primary ASR system designed for Broadcast news (BN), using as the secondary classifier an ASR system designed for conversational telephone speech (CTS); i.e. only un-transcribed speech from the Lecture domain is assumed to be given. Details of the experimental setup are described in Section 4, and results follow in Sections 5 and 6.

2. Theoretical Formulation

Theoretical justification for the proposed method comes from classification literature, via generalization error bounds on classifiers using randomization. We present the main idea here, relegating details to [6], where we have extended previous work on using unlabeled data in a co-validation setting [7], semi-supervised error bounds estimates [8], and structural risk minimization [9].

Let the *labeled* data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ comprise n i.i.d. samples of $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with common distribution $P_{X,Y}$ and let the *unlabeled* data $\{x_{n+1}, \ldots, x_{n+m}\}$ be m i.i.d. samples of X, whose common distribution P_X is the $\mathcal{X}$-marginal of $P_{X,Y}$. In our case, X is a sample of speech and Y is its orthographic transcript. A classifier is a mapping

$g : \mathcal{X} \to \mathcal{Y}$, and its generalization error is

$$\mathrm{e}(g) = P_{X,Y}[g(X) \neq Y]. \quad (1)$$

In our case, g will be an ASR system.

A *randomized* classifier $f : \mathcal{X} \to \mathcal{Y}$ is a random variable taking values in $\mathcal{Y}$. The definition (1) of generalization error extends easily to randomized classifiers by replacing $P_{X,Y}$ with $P_{X,Y,f}$, the joint probability of X, Y *and* the randomness in f.

Given two classifiers f and g, their *disagreement* is simply

$$d(g, f) = P_X[g(X) \neq f(X)], \quad (2)$$

which is approximated very well, when a large amount of unlabeled data are available ($m \gg 1$), by its empirical value

$$\hat{d}(g, f) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}[g(x_{n+j}) \neq f(x_{n+j})], \quad (3)$$

where $\mathbb{I}[\cdot]$ is the indicator function. Next, if a randomized classifier f^* is an *oracle*, i.e. $P_{X,Y,f^*}[f^*(X) = Y] = 1$, and g is any other (deterministic) classifier, then

$$d(g, f) = P_{X,Y,f^*}[g(X) \neq f^*(X)] = \mathrm{e}(g). \quad (4)$$

Therefore, if we are required to choose between two classifiers g_1 and g_2, and we have access to an *oracle* f^* along with some unlabeled data, a choice based on the empirical disagreement $\hat{d}(g_1, f^*)$ v/s $\hat{d}(g_2, f^*)$ on a held out data set is easily justified. In case of ASR, f^* is a human transcriber, making $\hat{d}(g, f^*)$ the empirical WER on the held out set.

The case when f is *not* an oracle but an error-prone labeler is considered in [6]. It is shown that under suitable conditions, such as assuming that $\mathrm{e}(f) \ll 1$ and that the randomness in f is such that given X, the disagreement event $\mathbb{I}[g_1(X) \neq f(X)]$ is uncorrelated with $\mathbb{I}[g_2(X) \neq f(X)]$, it continues to hold that

$$d(g_1, f) < d(g_2, f) \implies e(g_1) < e(g_2). \quad (5)$$

Since both the disagreements on the left are well estimated by their empirical values, one is justified in choosing between g_1 and g_2 based on a comparison of $\hat{d}(g_1, f)$ v/s $\hat{d}(g_2, f)$.

We therefore hypothesize that one may use the output f of an(other) ASR system in lieu of manual transcripts f^* to choose between two ASR systems g_1 and g_2, by trying to engineer the errors in the *sloppy reference* $f(X)$ to be uncorrelated with those of g_1 and g_2. This is the hypothesis investigated here.

3. Experimental Design

State of the art ASR systems are domain-specific. A small global change in acoustic conditions may cause all acoustic model probabilities to go down significantly while still maintaining some discrimination between acoustic-phonetic classes, or a modest shift in topic may cause many language model probabilities to go down while still providing useful relative distinctions between content words. In such situations, adjusting the scaling parameter that combines the acoustic and language models often provides significant relief. The LM scaling factor is traditionally tuned by minimizing the empirical WER on a held out set. But we will use the idea developed in Section 2 above to choose it in an unsupervised manner.

For a parametric family $\mathcal{G}^{\mathrm{BN}} = \{g_\lambda\}$ of ASR systems, whose acoustic and language models are trained on Broadcast news (BN) speech and text, we will investigate the problem of finding the LM scaling factor λ that minimizes WER on lecture speech X from the MIT Open Courseware corpus. i.e., we seek

$$\lambda^* = \arg\min_\lambda \mathrm{e}(g_\lambda) = \arg\min_\lambda P_{\mathrm{MIT}}(g_\lambda(X) \neq Y), \quad (6)$$

where Y denotes the manual transcript of X, and P_{MIT} the joint distribution of speech and transcripts in the MIT Lectures.

To estimate λ^* in an unsupervised manner, we take advantage of a family of ASR systems $\mathcal{F}^{\mathrm{CTS}} = \{f_\theta\}$, whose acoustic and language models are trained on conversational telephone speech (CTS) and transcripts, where θ again represents the LM scaling factor used for recognition.

In the first experiment in Section 5, we use the LM scaling factor inherited from the CTS corpus, i.e.

$$\theta_0 = \arg\min_\theta \mathrm{e}(f_\theta) = \arg\min_\theta P_{\mathrm{CTS}}(f_\theta(X) \neq Y), \quad (7)$$

to generate *sloppy references* for the MIT Lectures using f_{θ_0}, and use them to tune the LM scale factor of the BN system for transcribing the MIT Lectures:

$$\lambda^{(1)} = \lambda^{(1)}(\theta_0) = \arg\min_\lambda P_{\mathrm{MIT}}(g_\lambda(X) \neq f_{\theta_0}(X)). \quad (8)$$

Next, it is argued in [6] that the lower the error rate of f_θ, the better is $\hat{d}(g_\lambda, f_\theta)$ as a surrogate for minimizing the error rate of g_λ. In the second experiment in Section 6, we therefore begin with the LM scaling factor inherited from the BN corpus, i.e.

$$\lambda^{(0)} = \arg\min_\lambda \mathrm{e}(g_\lambda) = \arg\min_\lambda P_{\mathrm{BN}}(g_\lambda(X) \neq Y), \quad (9)$$

to generate *sloppy references* for the MIT Lectures using $g_{\lambda^{(0)}}$, and use them to tune the LM scale factor of the CTS system for transcribing the MIT Lectures:

$$\theta_1 = \theta_1(\lambda^{(0)}) = \arg\min_\theta P_{\mathrm{MIT}}(f_\theta(X) \neq g_{\lambda^{(0)}}(X)). \quad (10)$$

We then repeat (8), but use the CTS system f_{θ_1} instead of f_{θ_0}:

$$\lambda^{(2)} = \lambda^{(2)}(\theta_1) = \arg\min_\lambda P_{\mathrm{MIT}}(g_\lambda(X) \neq f_{\theta_1}(X)). \quad (11)$$

We compare the WERs $\mathrm{e}(g_{\lambda^*})$ with $\mathrm{e}(g_{\lambda^{(1)}})$ and $\mathrm{e}(g_{\lambda^{(2)}})$.

4. Experimental Setup

The CTS and BN ASR systems, $\mathcal{F}^{\mathrm{CTS}}$ and $\mathcal{G}^{\mathrm{BN}}$ above, are state of the art systems, trained using IBM's Attila tools [3].

The CTS system was taken 'as-is' from IBM's entry in the RT-04F evaluation. That system was discriminatively trained using MPE and fMPE with a recognition lexicon of 30.5K words. The acoustic models were trained using data from 5 sources: Fisher parts 1-7, Switchboard-1, BBN/CTRAN Switchboard-2, Switchboard Cellular, and Callhome English sampled at 8KHz.

The speaker-independent BN system was trained on 430 hours of audio from the 1996 English Broadcast News Speech corpus, the 1997 English Broadcast News Speech corpus, and the TDT4 Multilingual Broadcast News Speech corpus. The speaker-adapted BN system was trained on 6,096 hours of audio that included the three sources listed above as well as the English portion of the EARS BN03 data. PLP features are mean-variance-normalized on a segment-by-segment basis (speaker-independent pass) and on a speaker-by-speaker basis (speaker-dependent pass). Both passes use LDA+MLLT transforms to project to a 40-dimensional recognition feature space. Both

systems have acoustic models with roughly 6,000 states and 250,000 Gaussians and are discriminatively trained using fMPE and MPE with backing off to MMI estimates in I-smoothing. Details can be found in [11].

The BN system language model was trained using 335M words from the following sources: 1996 CSR Hub4 language model data, EARS BN03 closed captions, GALE Phase 2 Distillation GNG Evaluation Supplemental Multilingual data, Hub4 acoustic model training transcripts, TDT4 closed captions, TDT4 newswire, GALE Broadcast Conversations and GALE Broadcast News. The recognition lexicon size was around 84K words.

The λ's and θ's range from 0.010 to 0.500; in Attila, the LM scaling factor λ (or θ) multiplies the acoustic log-likelihood.

The MIT Lecture data [12] contains 20 Computer Science lectures, with a total of 7,346 utterances and 176,620 words from roughly 21 hours given by two speakers (each with 10 lectures). For the 84k BN vocabulary, the OOV rate of the MIT Lectures is 3.3%, while for the 40k CTS vocabulary, it is 3%. In order to use the CTS system, the lecture speech is downsampled to 8KHz.

5. Results and Analyses

Decoding the MIT Lectures with an out-of-the-box CTS system f_{θ_0} yields a WER of 37.01%. In the notation developed above,

$$\mathrm{e}(f_{\theta_0}) = P_{\mathrm{MIT}}(f_{\theta_0}(X) \neq Y) \approx 0.3701.$$

Figure 1 illustrates, for a range of λ-values, the disagreement $\hat{d}(g_\lambda, f_{\theta_0})$ between the parameterized BN systems and the default CTS system, as well as the true WER $P_{\mathrm{MIT}}(g_\lambda(X) \neq Y)$ of the BN system on MIT Lectures. It is clear that $\hat{d}(g_\lambda, f_{\theta_0})$ by itself is not a good estimate of $\mathrm{e}(g_\lambda)$, since $f_{\theta_0}(X)$ itself has a 37% WER. But it is also clear that the value $\lambda^{(1)}$ at which $\hat{d}(g_\lambda, f_{\theta_0})$ attains its minimum is fairly close to the value λ^* at which true WER is minimized.

Since $\hat{d}(g_\lambda, f_{\theta_0})$ is an unreliable estimate of true WER, it is reasonable to round it off to, say, the nearest 1%. Doing so results in a range of values $[0.040, 0.055]$—instead of a single value—for $\lambda^{(1)}$. Table 1 lists the true WER in this range. Note from Table 1 that choosing $\lambda^{(1)}$ strictly according to (8) results in a true WER of 25.06%. The range also happens to contain $\lambda^* = 0.055$, at which the true WER is 24.45%.

The WER of 24.45% was obtained by optimizing λ on the *test* set. To assess the WER performance of any $\lambda^{(1)}$ chosen from Table 1 or according to (8), without looking at the true labels on the test set, we also conducted the standard exercise of optimizing the BN system g_λ on 1 hour of held out MIT Lectures speech with reference transcripts. About 30 minutes of speech from each of the two test speakers was used. This yielded a $\lambda^{\mathrm{sup}} = 0.060$, with a corresponding WER of 24.50%.

It is clear from the upper half of Table 2 that while $\lambda^{(1)}$ is about as close to λ^* as the estimate λ^{sup} from an hour of transcribed speech, it does not improve over the default setting $\lambda^{(0)}$ for the BN system. Speculating that this is due to the high WER of the sloppy reference $f_{\theta_0}(X)$, we investigated the iterated estimate $\lambda^{(2)}$ of (11), and report results in Section 6.

But before doing so, we examine further the correlation between the disagreements between f_θ and g_λ, and errors committed by either one of them. We first align the sloppy reference $f_{\theta_0}(X)$ with the correct transcript of the MIT Lectures and mark all words that are transcribed in error. We then align the ASR system output $g_\lambda(X)$ with the correct transcript and again mark all words that are in error. Finally, we align the ASR system output with the sloppy reference. This enables us to determine the fraction of words that both systems recognized correctly (60.1%), the fraction "deemed" errors by the sloppy reference that were *false alarms* (16.1%), etc. These statistics are presented in Table 3.

We note further that of the 18.8% words that both systems recognized incorrectly (cf. Table 3), the two agreed on their incorrect answer, constituting *false negatives*, in 48.4% cases.

These correlations, however, may not be fatal to our method; what need to be uncorrelated are the conditional disagreements, *given* X, of two different ASR system outputs $g_\lambda(X)$ and $g_{\lambda'}(X)$ with the sloppy reference $f_\theta(X)$.

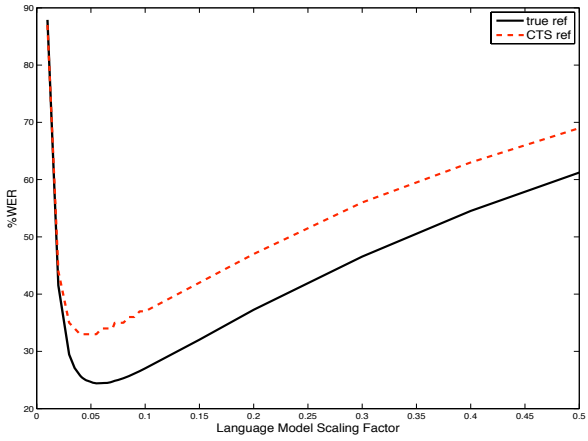

Figure 1: Variation of the true WER (solid) and WER w.r.t the sloppy reference (dashed) as a function of the LM scaling factor.

	λ	$\hat{d}(g_\lambda, f_{\theta_0})$	WER
	0.040	36%	25.81%
	0.041	36%	25.60%
	0.042	36%	25.44%
	0.043	36%	25.31%
	0.044	36%	25.19%
$\lambda^{(1)}$	**0.045**	36%	**25.06%**
	0.046	36%	24.96%
$\lambda^{(0)}$	0.047	36%	24.90%
	0.048	36%	24.83%
	0.049	36%	24.77%
	0.050	36%	24.69%
	0.051	36%	24.62%
	0.052	36%	24.55%
	0.053	36%	24.53%
λ^*	**0.055**	36%	**24.45%**

Table 1: The range of λ's that [approximately] minimize disagreement between the ASR output and a sloppy reference, and the true WER for those λ's. $\lambda^{(1)}$: minimizer of (exact) disagreement; λ^*: minimizer of true WER; $\lambda^{(0)}$: system default.

6. Iterative Estimation

Following the experimental design described in (10), we used the default setting $\lambda^{(0)}$ for the BN system to transcribe the MIT Lectures and used the resulting transcript as a sloppy reference to obtain the LM scaling factor θ_1 for the CTS system. The true

Models	LM Scaling	Value	WER
BN: g_λ	$\lambda^{(1)}$	0.045	25.1%
BN: g_λ	$\lambda^{(0)}$	0.047	24.9%
BN: g_λ	λ^*	0.055	24.5%
BN: g_λ	λ^{sup}	0.060	24.5%
CTS: f_θ	θ_0	0.060	37.1%
CTS: f_θ	θ_1	0.078	35.8%
CTS: f_θ	θ_*	0.085	35.8%
BN: g_λ	$\lambda^{(2)}$	0.052	24.6%

Table 2: WER on the MIT Lectures for various ASR systems.

		Sloppy Ref. (CTS)		
		Correct	In error	
ASR Output (BN)	Correct	60.2%	16.1%	76.3%
	In error	4.9%	18.8%	23.7%
		65.1%	34.9%	

Table 3: Joint distribution of correct/erroneous words in the BN and CTS system outputs shows that their errors are correlated.

WERs for the default (θ_0), tuned (θ_1) and optimal (θ_*) CTS systems are shown in the middle third of Table 2.

We then used the CTS system f_{θ_1} to produce a transcript of the *same* MIT Lecture data, and used this transcript as the sloppy reference for tuning the LM scaling factor for the BN system, as specified in (11). Finally, the resulting LM scaling factor $\lambda^{(2)}$ was used to transcribe the lectures one again. This iterative process is illustrated via a sketch in Figure 2, and the true WER of the re-estimated BN system $g_{\lambda^{(2)}}$ is shown in the last row of Table 2.

Note from Table 2 that the small improvement in the WER of the CTS system is optimal, and using the transcript of this CTS system as the sloppy reference improves the BN system to the point where the difference between *supervised* and *unsupervised* adjustment of the LM scaling factor is not significant any more. This lends further support to the hypothesis of Section 2.

7. Conclusions

In this paper we demonstrate an effective method for using unlabeled data to adjust the language model scaling factor in an ASR system to a new domain. We show that an imperfect yet readily available transcript generated by another system can serve as a *sloppy reference*. Despite many errors in the automatic transcript ($\sim$ 35% WER), we can still perform model selection (of the scaling factor) based on minimizing the error-rate (disagreement) w.r.t. this sloppy reference in a state-of-the-art system. Our experiments indicate that estimating the scaling factor when changing domains from Broadcast News to MIT Lectures can be done without the use of labeled "in-domain" data.

The quality of the sloppy reference is shown to play a role in its ability to serve as a proxy for the true reference, and further analysis is needed to shed light on when the technique breaks down. Furthermore it remains to be seen how the technique changes with the inherent difficulty of the problem. However, for this case, the simple method results in finding a nearly optimal scaling factor *without* requiring any manual transcripts.

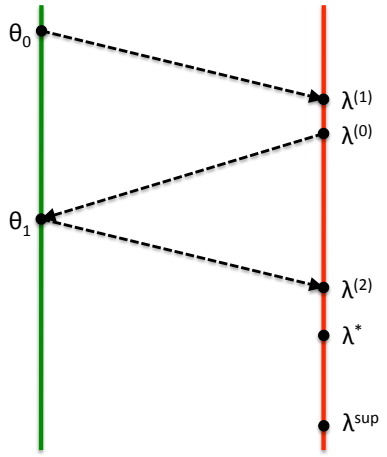

Figure 2: Instead of (i) obtaining $\lambda^{(1)}$ from a transcript produced by the default CTS setting θ_0, (ii) first obtain an improved θ_1 from a transcript produced by the default BN setting $\lambda^{(0)}$, and (iii) obtain $\lambda^{(2)}$ from a transcript produced by the improved CTS setting θ_1. λ^* is the best possible (oracle) setting, while λ^{sup} results from supervised estimation of the scaling factor.

8. Acknowledgements

We gratefully acknowledge the assistance of colleagues at IBM Research in the use of their LVCSR system, Attila [3].

9. References

[1] Lin Lawrence Chase, "Error-Responsive Feedback Mechanisms for Speech Recognizers", Tech. Report CMU-RI-TR-97-18, 1997

[2] Dimitra Vergyri, "Integration of Multiple Knowledge Sources in Speech Recognition using Minimum Error Training", PhD Dissertation, Johns Hopkins University, 2000.

[3] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig., "The IBM 2004 Conversational Telephony System for Rich Transcription", In *Proceedings of ICASSP*, 2005.

[4] Rion Snow, Brendan OConnor, Daniel Jurafsky, and Andrew Y. Ng, "Cheap and Fast - But is it Good? Evaluating Nonexpert Annotations for Machine Learning Tasks", In *Proceedings of EMNLP*, 2008.

[5] J. Ma and R. Schwartz, "Unsupervised versus Supervised Training of Acoustic Models", *in Proceedings of Interspeech*, 2008.

[6] Christopher M. White, "Semi-supervised model selection using likelihood ratios and disagreement on unlabeled data with applications in speech recognition", PhD dissertation, Johns Hopkins University, 2009.

[7] Madani, O., Pennock, D.M., and Flake, G.W., "Co-Validation: Using Model Disagreement on Unlabeled Data to Validate Classification Algorithms", *in Proceedings of NIPS*, 2004.

[8] Matti Kaariainen, "Generalization Error Bounds Using Unlabeled Data", *In COLT*,pp. 127-142, 2005.

[9] V. N. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, 1998.

[10] Avrim Blum, and Tom Mitchell, "Combining Labeled and Unlabeled Data with Co-Training", In *Proceedings of COLT*, 1998.

[11] S. Chen, B. Kingsbury, L. Mangu, D. Povey, G. Saon, H. Soltau and G. Zweig, "Advances in Speech Transcription at IBM under the DARPA EARS Program", *IEEE Transactions on Audio, Speech and Language Processing*, 2006.

[12] J. Glass, T. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent Progress in MIT Spoken Lecture Processing Project", *Proceedings of Interspeech*, 2007.