A Geometric Interpretation of Non-Target-Normalized Maximum Cross-channel Correlation for Vocal Activity Detection in Meetings

Kornel Laskowski
interACT, Universität Karlsruhe
Karlsruhe, Germany
kornel@ira.uka.de

Tanja Schultz
interACT, Carnegie Mellon University
Pittsburgh PA, USA
tanja@cs.cmu.edu

Abstract

Vocal activity detection is an important technology for both automatic speech recognition and automatic speech understanding. In meetings, standard vocal activity detection algorithms have been shown to be ineffective, because participants typically vocalize for only a fraction of the recorded time and because, while they are not vocalizing, their channels are frequently dominated by crosstalk from other participants. In the present work, we review a particular type of normalization of maximum cross-channel correlation, a feature recently introduced to address the crosstalk problem. We derive a plausible geometric interpretation and show how the frame size affects performance.

1 Introduction

Vocal activity detection (VAD) is an important technology for any application with an automatic speech recognition (ASR) front end. In meetings, participants typically vocalize for only a fraction of the recorded time. Their temporally contiguous contributions should be identified prior to ASR in order to leverage speaker adaptation schemes and language model constraints, and to associate recognized output with specific speakers (who said what). Segmentation into such contributions is informed primarily by VAD on a frame-by-frame basis.

Individual head-mounted microphone (IHM) recordings of meetings present a particular challenge for VAD, due to crosstalk from other participants. Most state-of-the-art VAD systems for meetings rely on decoding in a binary speech/non-speech space, assuming independence among participants, but are increasingly relying on features specifically designed to address the crosstalk issue (Wrigley et al., 2005).

A feature which has attracted attention since its use in VAD post-processing in (Pfau et al., 2001) is the maximum cross-channel correlation (XC), $\max_\tau \phi_{jk}(\tau)$, between channels j and k, where τ is the lag. When designing features descriptive of the kth channel, XC is frequently normalized by the energy in the target[1] channel k (Wrigley et al., 2003). Alternately, XC can be normalized by the energy in the non-target channel j (Laskowski et al., 2004), a normalization which we refer to here as NT-Norm, extending the Norm and S-Norm naming conventions in (Wrigley et al., 2005). Table 1 shows several types of normalizations which have been explored.

Normalization of XC		Mean	Min	Max
(none)	$\max_{j\neq k} \phi_{jk}(\tau)$	**[2][4]**	[2][4]	[2][4]
Norm	$\frac{\max_{j\neq k} \phi_{jk}(\tau)}{\phi_{kk}(0)}$	**[2][4]**	[2][4]	**[2][4]**
S-Norm	$\frac{\max_{j\neq k} \phi_{jk}(\tau)}{\sqrt{\phi_{jj}(0)\phi_{kk}(0)}}$	[2]**[4]**[5]	[2][4]	[1][2]**[4]**
NT-Norm	$\frac{\max_{j\neq k} \phi_{jk}(\tau)}{\phi_{jj}(0)}$	[3]	[6]	[6]

Table 1: Normalizations and statistics of cross-channel correlation features to describe channel k. In [1], a median-smoothed version was used in post-processing. In [3], the sum (JMXC) was used instead of the mean. In [5], cross-correlation was computed over samples and features. In [6], the minimum and the maximum were jointly referred to as NMXC. References in bold depict features selected by an automatic feature selection algorithm in [2] and [4]. (1:(Pfau et al., 2001), 2:(Wrigley et al., 2003), 3:(Laskowski et al., 2004), 4:(Wrigley et al., 2005), 5:(Huang, 2005), 6:(Boakye and Stolcke, 2006))

[1]The target/non-target terms are due to (Boakye and Stolcke, 2006).

Proceedings of NAACL HLT 2007, Companion Volume, pages 89–92,
Rochester, NY, April 2007.

The present work revisits NT-Norm normalization, which has been successfully used in a threshold detector (Laskowski et al., 2004), in automatic initial label assignment (Laskowski and Schultz, 2006), and as part of a two-state decoder feature vector (Boakye and Stolcke, 2006). Our main contribution is a geometric interpretation of NT-Norm XC, in Section 2. We also describe, in Section 3, several contrastive experiments, and discuss the results in Section 4.

2 Geometric Interpretation

We propose an interpretable geometric approximation to NT-Norm XC for channel k,

$$\xi_{k,j} \;=\; \frac{\max_\tau \phi_{jk}(\tau)}{\phi_{jj}}\,, \quad \forall j{\neq}k \tag{1}$$

We assume the simplified response in the kth IHM microphone at a distance d_k from a single point source $s(t)$ to be

$$m_k(t) \;\doteq\; A_k\left(\frac{1}{d_k}\,s\left(t-\frac{d_k}{c}\right)+\eta_k(t)\right), \tag{2}$$

where c, A_k and $\eta_k(t)$ are the speed of sound, the gain of microphone k, and source-uncorrelated noise at microphone k, respectively. Cross-channel correlation is approximated over a frame of size Ω by

$$\phi_{jk}(\tau) \;=\; \int_\Omega \frac{A_j A_k}{d_j d_k} s(t)\, s(t-\tau)\, dt\,, \tag{3}$$

where $\tau \equiv (d_j - d_k)/c$. Letting $\mathcal{P}_s \equiv \int_\Omega s^2(t)\,dt$ and $\mathcal{P}_{\eta_k} \equiv \int_\Omega \eta_k^2(t)\,dt$,

$$\phi_{jj}(0) \;=\; A_j^2\left(\frac{1}{d_j^2}\,\mathcal{P}_s + \mathcal{P}_{\eta_j}\right)\,, \tag{4}$$

$$\max_\tau \phi_{jk}(\tau) \;=\; \frac{A_j A_k}{d_j d_k}\mathcal{P}_s\,, \tag{5}$$

respectively, as the maximum of $\phi_{jk}(\tau)$ occurs at $\tau^* = (d_k - d_j)/c$. In consequence,

$$\frac{\max_\tau \phi_{jk}(\tau)}{\phi_{jj}(0)} \;\approx\; \frac{d_j}{d_k}\,, \tag{6}$$

provided that

$$\frac{A_k}{A_j}\left[1-\frac{\mathcal{P}_{\eta_j}}{\frac{1}{d_j^2}\mathcal{P}_s+\mathcal{P}_{\eta_j}}\right] \;\approx\; 1\,, \tag{7}$$

i.e., under assumptions of similar microphone gains, a non-negligible farfield signal-to-noise ratio at each microphone, and the simplifications embodied in Equation 2, NT-Norm XC approximates the relative distances of 2 microphones to the single point source $s(t)$. We stress that this approximation requires no side knowledge about the true positions of the participants or of their microphones.

Importantly, this interpretation is valid only if τ^* lies within the integration window Ω in Equation 3. In (Boakye and Stolcke, 2006), the authors showed that when the analysis window is 25 ms, the NMXC feature is not as robust as frame-level energy flooring followed by cross-channel normalization (NLED).

3 Experimental Setup

3.1 VAD and ASR Systems

Our multispeaker VAD system, shown in Figure 1, was introduced in (Laskowski and Schultz, 2006). Rather than detecting the 2-state speech ($\mathcal{V}$) vs. non-speech ($\mathcal{N}$) activity of each partipant independently, the system implements a Viterbi search for the best path through a 2^K-state vocal interaction space, where K is the number of participants. Segmentation consists of three passes: initial label assignment (ILA), described in the next subsection, for acoustic model training; simultaneous multi-participant Viterbi decoding; and smoothing to produce segments for ASR. In the current work, during decoding, we limit the maximum number of simultaneously vocalizing participants to 3.

This system is an improved version of that fielded in the NIST Rich Transcription 2006 Meeting Recognition evaluation (RT06s)[2], to produce automatic segmentation in the IHM condition on conference meetings. The ASR system which we use in this paper is as described in (Fügen et al., 2007).

3.2 Unsupervised ILA

For unsupervised labeling of the test audio, prior to acoustic model training, we employ the criterion

$$\tilde{\mathbf{q}}[k] \;=\; \begin{cases} \mathcal{V} & \text{if } \sum_{j\neq k} \log\left(\frac{\max_\tau \phi_{jk}(\tau)}{\phi_{jj}(0)}\right) > 0 \\ \mathcal{N} & \text{otherwise} \quad . \end{cases} \tag{8}$$

Assuming equality in Equation 6, this corresponds to *declaring a participant as vocalizing when the distance between the location of the dominant sound source and that participant's microphone is smaller than the geometric mean of the distances from the source to the remaining microphones*, ie. when

$$\sqrt[K-1]{\prod_{j\neq k} d_j} \;>\; d_k \tag{9}$$

[2] `http://www.nist.gov/speech/tests/rt/`

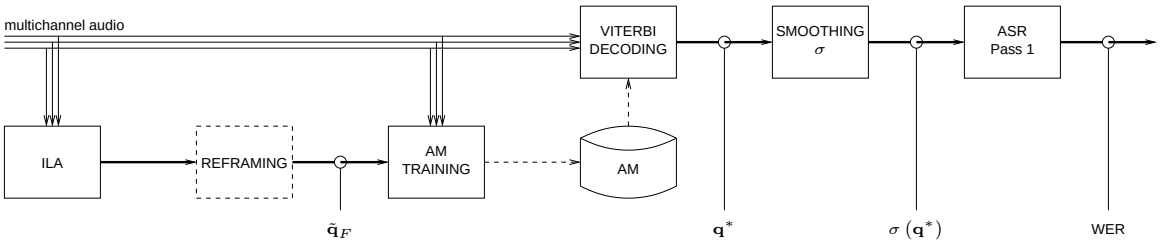

Figure 1: VAD system architecture, with 4 error measurement points. Symbols as in the text.

We refer to this algorithm as ILAave. For contrast we also consider ILAmin, with the sum in Equation 8 replaced by the minimum over $j \neq k$. This corresponds to *declaring a participant as vocalizing when the distance between the location of the dominant sound source and that participant's microphone is smaller than the distance from the source to any other microphone*. We do not consider ILAmax, whose interpretation in light of Equation 6 is not useful.

3.3 Data

The data used in the described experiments consist of two datasets from the NIST RT-05s and RT-06s evaluations. The data which had been used for VAD system improvement, `rt05s_eval*`, is the complete `rt05s_eval` set less one meeting, `NIST_20050412-1303`. This meeting was excluded as it contains a participant without a microphone, a condition known a priori to be absent in `rt06s_eval`; we use the latter in its entirety.

3.4 Description of Experiments

The experiments we present aim to compare ILAave and ILAmin, and to show how the size of the integration window, Ω, affects system performance. As our VAD decoder operates at a frame size of 100ms, we introduce a reframing step between the ILA component and both AM training and decoding; see Figure 1. $\mathcal{V}$ is assigned to each 100ms frame if 50% or more of the frame duration is assigned $\mathcal{V}$ by ILA; otherwise, the 100ms frame is assigned an $\mathcal{N}$ label.

We measure performance in four locations within the combined VAD+ASR system architecture, also shown in Figure 1. We compute a VAD frame error just after reframing ($\tilde{\mathbf{q}}_F$), just after decoding ($\mathbf{q}^*$), and just after smoothing ($\sigma(\mathbf{q}^*)$). This error is the sum of the miss rate (MS), and the false alarm rate excluding intervals of all-participant silence (FAX), computed against unsmoothed word-level forced alignment references. We use this metric for comparative purposes only, across the various measurement points. We also use first-pass ASR word error rates (WERs), after lattice rescoring, as a final measure of performance impact.

We evaluate, over a range of ILA frame sizes, the performance of ILAave(3), with a maximum number of simultaneously vocalizing participants of 3, and for the contrastive ILAmin. We note that ILAmin is capable of declaring at most one microphone at a time as being worn by a current speaker. As a result, construction of acoustic models for overlapped vocal activity states, described in (Laskowski and Schultz, 2006), results in states of at most 2 simultaneously vocalizing participants. We therefore refer to ILAmin as ILAmin(2), and additionally consider ILAave(2), in which states with 3 simultaneously vocalizing participants are removed.

4 Results and Discussion

We show the results of our experiments in Table 2. First-pass WERs, using reference segmentation (`.stm`), vary by 1.3% absolute (abs) between `rt05s_eval` and `rt06s_eval`. We also note that removing the one meeting with a participant without a microphone reduces the `rt05s_eval` manual segmentation WER by 1.7% abs. WERs obtained with automatic segmentation should be compared to the manual segmentation WERs for each set.

As the $\tilde{\mathbf{q}}_F$ columns shows, ILAmin(2) entails significantly more VAD errors than ILAave. Notably, although we do not show the breakdown, ILAmin(2) is characterized by fewer false alarms, but misses much more speech than ILAave(2). This is due in part to its inability to identify simultaneous talkers. However, following acoustic model training and use ($\mathbf{q}^*$), the VAD error rates between the two algorithms are approximately equal.

In studying the WERs for each ILA algorithm independently, the variation across ILA frame sizes in the range 25–100 ms can be significant: for example, it is 1.2% abs for ILAmin(2) on `rt06s_eval`, compared to the difference with manual segmentation of 3.1% abs. Error curves, as a function of ILA frame size, are predominantly shallow parabolas, except at 75 ms (notably for ILAmin(2) at $\tilde{\mathbf{q}}_F$); we believe that

ILA	Ω	VAD, rt05s			WER, 1st pass		
		$\tilde{\mathbf{q}}_F$	$\mathbf{q}^*$	$\sigma(\mathbf{q}^*)$	05	05*	06
ave3	100	**31.3**	16.7	16.0	39.0	34.1	39.6
	75	33.6	**16.6**	**15.9**	38.9	34.1	39.9
	50	35.2	16.7	16.0	**38.8**	**34.0**	**39.3**
	25	36.8	17.3	16.3	39.6	34.2	39.7
ave2	100	**31.3**	15.8	15.2	37.8	34.4	39.7
	75	33.6	**15.6**	**15.0**	37.9	34.4	39.6
	50	35.2	15.8	15.2	**37.6**	**34.3**	**39.3**
	25	36.8	16.4	15.6	38.1	**34.3**	39.5
min2	100	**43.4**	15.8	14.7	38.2	35.2	**39.3**
	75	51.9	**15.6**	**14.6**	38.1	35.2	**39.3**
	50	47.1	15.7	**14.6**	**37.9**	**35.1**	40.1
	25	47.7	16.2	14.9	38.1	35.4	40.5
refs		9.5	9.5	9.5	36.1	34.4	37.4

Table 2: VAD errors, measured at three points in our system, and first-pass WERs for `rt05s_eval` (`05`), as well as first-pass WERs for `rt05s_eval*` (`05*`) and `rt06s_eval` (`06`). Results are shown for 3 contrastive VAD systems (ILAave(3), ILAave(2) and ILAmin(2)), and 4 ILA frame sizes (100ms, 75ms, 50ms, and 25ms).

this is because 75 ms does not divide evenly into the decoder frame size of 100 ms, causing more deletions across the reframing step than for other ILA frame sizes. Error minima appear for an ILA frame size somewhere between 50 ms and 75 ms, for both ASR and post-decoding VAD errors.

Although (Pfau et al., 2001) considered a maximum lag of 250 samples (15.6ms, or 5m at the speed of sound), their computation of S-Norm XC used a rectangular window. Here, as in (Laskowski and Schultz, 2006) and (Boakye and Stolcke, 2006), we use a Hamming window. Our results suggest that a large, broadly tapered window is important for Equation 6 to hold.

The table also shows that for datasets without uninstrumented participants, `rt05s_eval*` and `rt06s_eval`, ILAmin(2) is outperformed by ILAave(2) by as much as 1.1% abs in WER, especially at small frame sizes. The difference for the full `rt05s_eval` dataset is smaller. The results also suggest that reducing the maximum degree of simultaneous vocalization from 3 to 2 during decoding is an effective means of reducing errors (ASR insertions, not shown) for uninstrumented participants.

5 Conclusions

We have derived a geometric approximation for a particular type of normalization of maximum cross-channel correlation, NT-Norm XC, recently introduced for multispeaker vocal activity detection. Our derivation suggests that it is effectively comparing the distance between each speaker's mouth and each microphone. This is novel, as geometry is most often inferred using the lag of the crosscorrelation maximum, rather than its amplitude.

Our experiments suggest that frame sizes of 50–75 ms lead to WERs which are lower than those for either 100 ms or 25 ms by as much as 1.2% abs; that ILAave outperforms ILAmin as an initial label assignment criterion; and that reducing the degree of simultaneous vocalization during decoding may address problems due to uninstrumented participants.

6 Acknowledgments

This work was partly supported by the European Union under the integrated project CHIL (IST-506909), Computers in the Human Interaction Loop.

References

K. Boakye and A. Stolcke. 2006. Improved Speech Activity Detection Using Cross-Channel Features for Recognition of Multiparty Meetings. *Proc. of INTERSPEECH*, Pittsburgh PA, USA, pp1962–1965.

C. Fügen, S. Ikbal, F. Kraft, K. Kumatani, K. Laskowski, J. McDonough, M. Ostendorf, S. Stüker, and M. Wölfel. 2007. The ISL RT-06S Speech-to-Text Evaluation System. *Proc. of MLMI*, Springer Lecture Notes in Computer Science **4299**, pp407–418.

Z. Huang and M. Harper. 2005. Speech Activity Detection on Multichannels of Meeting Recordings. *Proc. of MLMI*, Springer Lecture Notes in Computer Science **3869**, pp415–427.

K. Laskowski, Q. Jin, and T. Schultz. 2004. Crosscorrelation-based Multispeaker Speech Activity Detection. *Proc. of INTERSPEECH*, Jeju Island, South Korea, pp973–976.

K. Laskowski and T. Schultz. 2006. Unsupervised Learning of Overlapped Speech Model Parameters for Multichannel Speech Activity Detection in Meetings. *Proc. of ICASSP*, Toulouse, France, I:993–996.

T. Pfau and D. Ellis and A. Stolcke. 2001. Multispeaker Speech Activity Detection for the ICSI Meeting Recorder. *Proc. of ASRU*, Madonna di Campiglio, Italy, pp107–110.

S. Wrigley, G. Brown, V. Wan, and S. Renals. 2003. Feature Selection for the Classification of Crosstalk in Multi-Channel Audio. *Proc. of EUROSPEECH*, Geneva, Switzerland, pp469–472.

S. Wrigley, G. Brown, V. Wan, and S. Renals. 2005. Speech and Crosstalk Detection in Multichannel Audio. *IEEE Trans. on Speech and Audio Processing*, **13**:1, pp84–91.