EXPLOITING A 'GAZE-LOMBARD EFFECT' TO IMPROVE ASR PERFORMANCE IN ACOUSTICALLY NOISY SETTINGS

Neil Cooke, Ao Shen, Martin Russell

School of Electronic and Electrical and Computer Engineering, University of Birmingham, Birmingham UK

ABSTRACT

Previous use of gaze (eye movement) to improve ASR performance involves shifting language model probability mass towards the subset of the vocabulary whose words are related to a person's visual attention. Motivated to improve Automatic Speech Recognition (ASR) performance in acoustically noisy settings by using information from gaze selectively, we propose a 'Selective Gaze-contingent ASR' (SGC-ASR). In modelling the relationship between gaze and speech conditioned on noise level - a 'gaze-Lombard effect' - simultaneous dynamic adaptation of acoustic models and the language model is achieved. Evaluation on a matched set of gaze and speech data recorded under a varying speech babble noise condition yields WER performance improvements. The work highlights the use of gaze information in dynamic model-based adaptation methods for noise robust ASR.

Index Terms— ASR, speech, acoustic noise, noise robust ASR. eye movement, Mutual information, Language Model adaptation, Acoustic Model adaptation, gaze, visual attention

1. INTRODUCTION

Gaze-contingent Automatic Speech Recognition (GC-ASR) uses information from gaze (eye movement) to improve performance. This involves the detection of deictic eye gestures i.e. visual attention (VA) towards visual referents. Performance improvements are achieved around the time of VA by dynamic Language Model (LM) adaptation - shifting LM probability mass away from words not related to the visual referent towards words which are. This could be desirable in acoustically noisy environments where speech is more difficult to detect; ASR difficulties arise primarily from deviations between training and field/test data resulting from source separability and spectral changes to speech (the 'Lombard effect' [1] [2]). Approaches to address the Lombard effect in the acoustic domain focus on feature enhancement and/or model parameter adaptation.

A GC-ASR function is increasingly afforded in human-computer interaction (HCI) systems. Technology affords the sensing of multiple human modalities and interactions are becoming more human-like e.g. robots and avatars. HCI systems are evolving from requiring people to actively and consciously direct gaze to display e.g. as a mouse replacement for people with physical disability, to systems that are aware of the user's visual environment and able to monitor natural gaze, placing no constraints a person's eye movements. Understanding speech communication in these scenarios requires consideration of the relationship of speech to other modalities and the use of information from other modalities in speech processing. A GC-ASR function makes a worthwhile contribution to these considerations.

Gaze has been ascribed several roles relating to a person's cognition and its role in human interaction [3]. If gaze is used as an information source for dynamic LM Adaptation in ASR in real-world settings, then its use should be selective. Furthermore, in the presence of acoustic noise, gaze behaviour and its relationship with speech may change, suggesting a potential for dynamic Acoustic Model (AM) adaptation by exploiting a 'gaze-Lombard effect'.

In this work a 'Selective Gaze-contingent Automatic Speech Recognition' (SGC-ASR) system is proposed. It uses information from gaze to improve its performance by both dynamic LM and AM adaptation. In section 2 related work is summarized. Section 3 describes the architecture and framework for the system. The implementation and evaluation is described in section 4, with results reported in section 5. Section 6 concludes.

2. RELATION TO PRIOR WORK

2.1. Using gaze to improve ASR

Dynamic LM Adaptation is the approach used in all current GC-ASR[4][5]. Refinements to use only VA information related to task have been proposed by exploiting differences in gaze before and during utterances[6] and estimating the relevance of a sequence of VA to task[7]. In these works, a cache-based LM is used where the cache contains sequence of VA with associated related words.

2.2. Novelty

Previous GC-ASR exclusively focus on dynamic LM adaptation. The effect of acoustic noise on performance of such systems was explored in our recent work [7]. The idea is ex-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

tended here by proposing a taxonomy of 'gaze roles' to discriminate between different gaze behaviours and to model the relationship between gaze and speech affording dynamic AM adaptation.

2.3. Explaining gaze behaviour

There are multiple explanations for gaze. Pioneering psychological studies linked gaze to cognitive processes [8] and social interaction [9]. Processes of cognition include scene perception, reading, object naming [10] [11] [12] and lexical processing [13]. Eliciting and attributing gaze to specific cognitive process is problematic because of the absence of a ground truth; cognition cannot be directly measured. Gaze can be more reliably interpreted in relation to VA during a system task or activity, communications with others and reaction to changes in the environment, due to a measurable ground truth i.e. identification of what is being looked at.

3. APPROACH

3.1. Taxonomy of gaze roles

Building on the work in [11], [14] and [15], a working (non-exhaustive) taxonomy for gaze in this study is proposed: *Cognition Roles* are gaze events relating to cognition which lack ground truth e.g. reading, scene perception, visual memory and object naming, sentence planning and psycholinguistic roles; *Visual Attention Roles* are gaze events relating to VA which have ground truth which include *Task-oriented Visual Attention(ToVA)* (VA events associated with tasks and activities assumed by the system), *Reactive Visual Attention* (RVA) (VA events elicited by changes to the environment) and *Social Visual Attention*(SVA) (gaze associated with social interaction and not considered further in this study).

3.2. Architecture

Figure 1 outlines the SGC-ASR. There are separate inference procedures for cognition and VA gaze roles. VA Inference affords LM Adaptation. Acoustic Noise Inference (via cognition roles) affords AM adaptation. Speech for the inference procedures is supplied from a baseline ASR system. Gaze is captured from an eye tracker which provides temporal and spatial gaze characteristics and the identification of the focus of VA.

3.3. Visual Attention Inference for LM Adaptation

Dynamic adaptation of an LM using information from gaze requires the identification of a subset of VA events - i.e. instances of viewing referents which have related words. Referring to section 3.1, these events are defined as Task-oriented VA (ToVA). An inference procedure is required to discriminate between ToVA and the other VA types. This is achieved

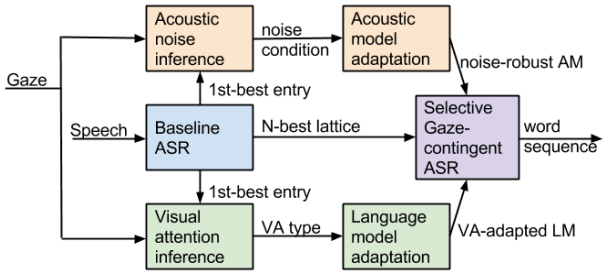

Fig. 1. *Architecture for the SGC-ASR system.*

by maximum likelihood supervised training of a Naive Bayes classifier with the input features standard gaze characteristics from VA events, namely fixation duration and saccade length. Ground truths for VA types are determined from rules based on their definitions e.g. a VA event of looking at a referent and speaking about it is labeled as ToVA whereas looking at an referent in response to another talking about it is labeled as RVA. An additional Task-independent Visual Attention (TiVA) class is used for all other VA events not covered by such rules. The trained classifier provides a score (or 'relevance') $\sigma_x \in [0, 1]$ for a VA event x which indicates its influence over LM Adaptation. The adapted LM word probability $P_a^t(w_i)$ at time t is determined by the weighted interpolation of a time invariant baseline n-gram LM $P_b(W)$ and an l length VA event cache-based LM $P_v^t(W)$ with word probabilities, $P_v^t(w_i)$, computed with the scores σ_x assigned to v_x: $P_v^t(w_i) = \frac{\sum_{\alpha} \sigma_y}{\sum_{l} \sigma_x}$ where α is the length of a subset of the VA cache with every VA event related to word w_i. More details of the VA cache-based LM using VA events and relevance scores is reported in [7].

3.4. Acoustic Noise Inference for Acoustic Model Adaptation

The dynamic adaptation of the ASR AM requires inference of the acoustic noise level. Features for this inference are the gaze characteristics fixation duration and saccade length and the relationship between gaze and speech. This relationship is captured with two features, each representing a cognition-oriented gaze role: Mediating Attention (MA) and Object Naming (ON), estimated over a temporal window by measures based on apriori knowledge. Figure 2 illustrates the approach. The top half shows the gaze-speech relationship for MA, and the bottom half for ON over the same period. Gaze and speech are modeled as a sequence of discrete-valued random variables that represent the 'information events' used to measure the presence of ON and MA. For gaze this is a VA on the visual referent (e.g. a shape) and the words spoken related to it (e.g. the name of the shape). A couple is defined for event pairs. Each couple has a strength r determined from the pair's temporal and semantic relationship. Referring to the

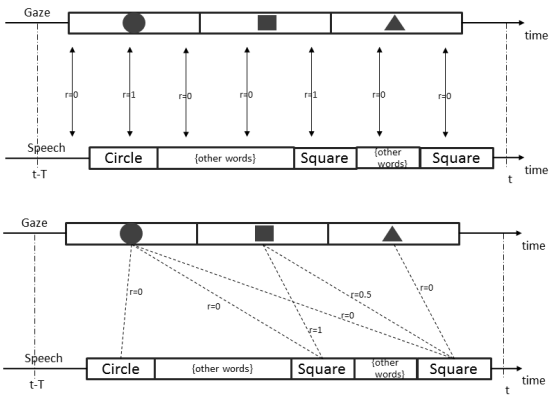

Fig. 2. *Measuring the gaze/speech relationship for gaze roles MA (top half) and ON (bottom half). r is the strength of the couples used to calculate Mutual Information for each role.*

figure, looking at the square prior to saying 'square' for MA (top half) has a strength of $r = 1$ if the gaze and speech overlap and $r = 0$ otherwise. Whereas for the ON role, a strength of $r = 0.5$ is calculated between non-overlapping events of looking at speaking the word 'square'. The prevalence of the gaze roles over a temporal window are estimated using Shannon's Mutual Information (MI) [16]. Let e_t and s_t be the random variables at time t for gaze and speech respectively. The MI at time t, $I(e_t; s_t)$, is a measure of the difference in entropy between the joint density $p(e_t, s_t)$ and the product of the marginal densities $p(e_t)$ and $p(s_t)$. An MI $> 0bits$ (assuming base 2 log) indicates the degree of presence of the gaze role. The joint and marginal densities from the coupled event pairs over a window of time T up to time t are estimated from the couples $r_{e,s}$: $p(e_t = e, s_t = s) = r_{e,s} \frac{N_{e,s}}{N_o}$ where $N_{e,s}$ is the number of couples between events e and s and N_o is the total number of couples observed. Within any temporal window T over which MI is calculated, there may be events that are observed only in another window e.g. other referents and words. Therefore, to preserve the axiomatic assumption of unit measure, the joint density for events seen in the data, but unseen as event couples in window T are uniformly estimated from the probability mass not assigned to the seen joint probabilities. Because the constituents of the joint density are observed events, the MI for different temporal windows becomes comparable regardless of its constituents. Marginal densities are calculated from the joint density.

4. EVALUATION

4.1. Data

The evaluation uses a dataset recorded for a 'put that there' [17] task. The user tells a wizard / instruction receiver to position a coloured shape on a map displayed on a computer screen, affording the user to use the cognition gaze role of ON and a VA gaze role of referring to the referents. The wizard selects and positions the object, the resulting dialogue affording the opportunity for the user to react to the wizards speech and display changes eliciting the users VA-role of RVA. The map displayed on the computer screen is augmented with the user's VA to afford the cognition role of MA - i.e. intention to guide the wizard. Seven participants take part. Amplified speech babble noise from noisex-92 [18] is added to the speech heard through headphones resulting in the task being undertaken under four acoustic noise conditions: no noise (' N0'), less than normal speech $43dB$ ('N1'), conversational speech $55dB$ ('N2') and outdoor commercial areas $64dB$ ('N3') . 100 tasks are recorded for each of the noise conditions. The user's gaze is captured using a head-mounted eye-tracker (SR Research Eyelink 2) capturing binocular eye position data at $500Hz$ and the corresponding fixation/saccade events. Participant's clean speech is recorded on separate audio channels at a sample rate of $44.1Khz$. Timestamped fixation events in the gaze data are assigned to their nearest visual focus, i.e. a color, shape or position on the map. Speech is time-aligned transcribed in two passes by human and forced alignment ASR system [19]. [7] gives further information regarding the data collected.

4.2. Baseline and dynamic LM Adaptation system

The baseline ASR was built using HTK [20] and trained on the WSJCAM0 corpus of British English with a dictionary of over 22000 pronunciations [21]. [5] gives further details. The baseline LM is constructed from the speech transcriptions of the data containing 1056 utterances and 3764 words. Bigrams with Witten-Bell smoothing[22] are used. Two AM sets are adapted to the data for speech in $N0$ and $N3$ conditions. Baseline AMs are adapted to the data using Maximum Likelihood Linear Regression (MLLR) [23] and Maximum A-Posteriori (MAP) adaptation [24]. Output for SGC-ASR using dynamic LM Adaptation is generated by rescoring baseline N-Best output for the visual cache-based LM as described in [7].

4.3. Tests

Three tests are conducted: *'gaze-Lombard effect'* - the changes to speech, gaze and their relationship (measured by MI) in acoustic noise are analyzed to motivate the use of features suitable for robust between-person inference of the acoustic noise condition; *Acoustic noise inference* - a discriminative classifier (Support Vector Machine (SVM)) infers the acoustic noise condition using the baseline speech, gaze and MI features; *ASR performance* - dynamic AM adaptation (i.e. selection of AM set as $N0$ or $N3$ given SVM output) and LM is compared against the baseline and the LM adaptation system reported in [7].

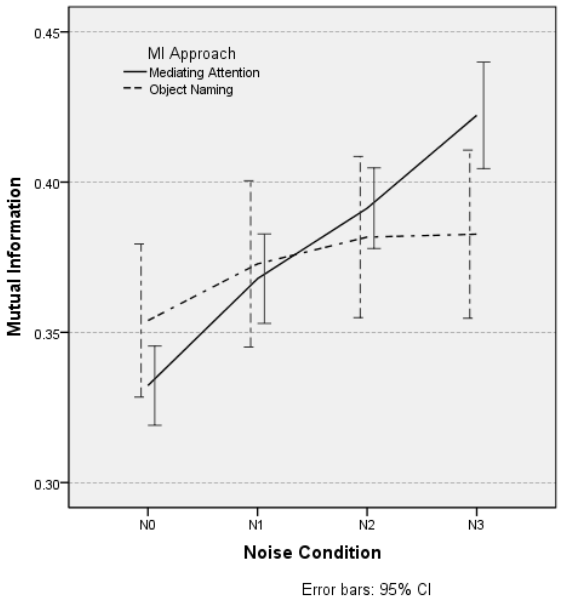

Fig. 3. *The MI values based on the gaze role of mediating attention(MA) and object naming(ON) respectively. A significant increase of MI value is observed for the role of MA as acoustic noise is increased in the environment*

5. RESULTS

5.1. 'Gaze-Lombard effect'

Across all 400 recorded tasks, as acoustic noise increases there is a significant increase in spectral power, $F0$, and a reduction in speech rate supporting previous researches of the Lombard effect. Fixation durations and saccade lengths significantly increase and decrease respectively, with changes depending on whether the user is speaking (mean duration increased 18.4% to $388ms$) or listening (mean duration increased 7.7% to $299ms$). Figure 3 shows that the MI measure for gaze role MA significantly increases with noise from $0.33bits$ to $0.42bits$. There is no significant increase for the gaze role ON MI measure. Changes in their relative values suggest a difference in cognition as acoustic noise increases.

5.2. Acoustic Noise Inference for dynamic LM adaptation

Two-class ($N0$ and $N3$) SVMs are trained with the radial basis function (RBF) kernel. Input features for the SVMs are speech, gaze and MI. A 10-fold cross-validation yields the classification results in Table 1. The MI feature set performs best with accuracy of 72.3% compared to gaze (51.6%) and speech (54.3%). Combinations of feature vectors for MI, gaze and speech perform no better to MI. An additional SVM discriminating the 7 participants reveals MI ore robust to between-person differences - 4% above chance accuracy, compared to the 18% and 29% for gaze and speech respectively.

Input Features	Acc	Precision	Recall	F-Measure
MI	0.723	0.728	0.723	0.717
Speech	0.516	0.487	0.516	0.464
Gaze	0.543	0.295	0.543	0.383

Table 1. *SVM classifier performance for $N0$ and $N3$. MI-based features infer the noise condition best.*

System	WER
Baseline	67.3
LM Adaptation	49.9
AM+LM Adaptation	**42.1**
LM +(perfect noise detection)	23.1*

Table 2. *ASR system performance comparisons demonstrating simultaneous dynamic LM and AM adaptation using gaze information improves performance.* *hypothetical*

5.3. SGC-ASR Performance

N-best list rescoring ($N = 100$) of ASR output is used to measure WER changes. Table 2 shows that the baseline WER of 67.28% is improved 17.4% by LM Adaptation. A further 7.8% improvement is gained from AM adaptation. For completeness, the hypothetical upper bound on performance assuming perfect noise detection, i.e. 100% accuracy for acoustic noise inference, yields a further 19% improvement.

6. SUMMARY

An SCG-ASR system employing dynamic language and acoustic model adaptation using gaze information selectively is described and evaluated on acoustically noisy speech. The inference of acoustic noise based on measuring changes in the gaze-speech relationship with information-theoretic mutual information demonstrates good accuracy and minimal between-person variation. The results suggest that a task-specific 'gaze-Lombard effect' may be exploited for noise-robust ASR performance by model adaptation. Further performance improvements by combining SGC-ASR with existing acoustic feature enhancement and model adaptation methods are feasible. The constrained user task setting for the evaluation invites further analysis of speech communication during other tasks where gaze is tracked in relation to the environment and other people.

7. REFERENCES

[1] E. Lombard, "Le signe de lelevation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, no. 101-119, pp. 25, 1911.

[2] H. Boril and J.H.L. Hansen, "Unsupervised equalization of lombard effect for speech recognition in noisy adverse environments," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1379–1393, 2010.

[3] ZM Griffin and K. Bock, "What the eyes say about speaking.," *Psychol Sci*, vol. 11, no. 4, pp. 274–9, 2000.

[4] S. Qu and J.Y. Chai, "An exploration of eye gaze in spoken language processing for multimodal conversational interfaces," *Proc. of North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*, pp. 284–291, 2007.

[5] N. Cooke and M. Russell, "Gaze-contingent automatic speech recognition," *IET signal processing*, vol. 2, no. 4, pp. 369–380, 2008.

[6] N.J. Cooke and M.J. Russell, "Cache-based language model adaptation using visual attention for asr in meeting scenarios," in *Proceedings of the 2009 international conference on Multimodal interfaces*. ACM, 2009, pp. 87–90.

[7] Shen A., N.J. Cooke, and M.J. Russell, "Selective use of gaze information to improve asr performance in noisy environments by cache-based class language model adaptation," in *Proc. Interspeech, Lyon, France, August 2013*, 2013.

[8] A.L. Yarbus, *Eye movements and vision*, Plenum press, 1967.

[9] A. Kendon, "Some functions of gaze-direction in social interaction.," *Acta psychologica*, vol. 26, no. 1, pp. 22–63, 1967.

[10] K. Rayner, "Eye movements and attention in reading, scene perception, and visual search," *The quarterly journal of experimental psychology*, vol. 62, no. 8, pp. 1457–1506, 2009.

[11] K. Rayner, "Eye movements in reading: Models and data," *Journal of eye movement research*, vol. 2, no. 5, pp. 1, 2009.

[12] K. Rayner, T.J. Smith, G.L. Malcolm, and J.M. Henderson, "Eye movements and visual encoding during scene perception," *Psychological Science*, vol. 20, no. 1, pp. 6, 2009.

[13] Z.M. Griffin, "Why look? Reasons for eye movements related to language production," *In M. Henderson and F. Ferreira Eds., The interface of language, vision, and action: Eye movements and the visual world*, pp. 213–247, 2004.

[14] V. Srinivasan and R. Murphy, "A survey of social gaze," in *Proceedings of the 6th international conference on Human-robot interaction*. ACM, 2011, pp. 253–254.

[15] S. Norris, *Analyzing multimodal interaction: A methodological framework*, Psychology Press, 2004.

[16] C.E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[17] C. Schmandt and E.A. Hulteen, "The intelligent voice-interactive interface," in *Proceedings of the 1982 conference on Human factors in computing systems*. ACM, 1982, pp. 363–366.

[18] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[19] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on Hidden Markov Models," *Speech Communication*, vol. 12, no. 4, pp. 357–370, 1993.

[20] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book*, vol. 2, Citeseer, 1997.

[21] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAMO: A British English speech corpus for large vocabulary continuous speech recognition," in *icassp*. IEEE, 1995, pp. 81–84.

[22] Ian H Witten and Timothy C Bell, "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression," *Information Theory, IEEE Transactions on*, vol. 37, no. 4, pp. 1085–1094, 1991.

[23] Christopher J Leggetter and PC Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[24] J-L Gauvain and Chin-Hui Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Speech and audio processing, ieee transactions on*, vol. 2, no. 2, pp. 291–298, 1994.