GAZE-CONTINGENT ASR FOR SPONTANEOUS, CONVERSATIONAL SPEECH: AN EVALUATION

Neil Cooke, Martin Russell

Multi-modal Interaction Laboratory, University of Birmingham, United Kingdom

ABSTRACT

There has been little work that attempts to improve the recognition of spontaneous, conversational speech by adding information from a loosely-coupled modality. This study investigated this idea by integrating information from gaze into an ASR system. A probabilistic framework for multimodal recognition was formalised and applied to the specific case of integrating gaze and speech. Gaze-contingent ASR systems were developed from a baseline ASR system by redistributing language model probability mass according to the visual attention. The best performing systems had similar Word Error Rates to the baseline ASR system and showed an increase in keyword spotting accuracy. The key finding was that performance improvements observed were due to increased recognition accuracy for words associated with the visual field but not the current focus of visual attention.

Index Terms— Speech recognition, Bayes procedures, Visual system, User interfaces

1. INTRODUCTION

Multimodal interfaces that are aware of the users attention are becoming more common and form a core part of work in multimodal human computer interaction [1]. The development of system architectures to handle multimodal dialogue including correcting recognition errors is a current research topic [2]. Accordingly, system functions that recognise must be realised; for speaking there is Automatic Speech Recognition (ASR) to recover word sequences; for deictic gestures such as eye movement or gaze, there is the recognition of attentive cues.

Motivated to achieve robust multimodal decoding functions in systems by using the information from one modality to improve recognition of another, we have undertaken a study that formalises a probabilistic multimodal recognition framework and have applied it to integrating visual attention with conversational speech. These 'Gaze-contingent' ASR systems aim to improve the recognition of the speech and, consequently, improve the recognition of the communicative intent.

For these systems, we assert that the communicative intent of the user manifests itself in speech through the use of particular words or grammatical structures, and in eye movement as looking at an individual visual focus or a sequence of visual foci (i.e. visual attention). Therefore, different communicative intents will correspond to different language models and sequences of visual foci. Accordingly, we believe that improved keyword spotting accuracy is an appropriate criterion for measuring the success of these systems in addition to the more popular Word Error Rates (WER), because it enables the word recognition performance to be measured for a subset of the recogniser's vocabulary; that is, the subset of the vocabulary related to the communicative intent.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 formalises recognition problem and outlines our gaze-contingent ASR system architecture. Section 4 describes the realisation of the gaze-contingent ASR system variants and their evaluation: How should it be implemented to realise recognition performance benefits? Section 5 presents some results. Section 6 answers this question and suggests future directions for this work.

2. RELATED WORK

In a recent comparable study, minor improvements in WER were reported when making an ASR system gaze-contingent, although the best performing system still had a high WER of 68.9% [3]. The gaze-contingent ASRs in this study differ because we are interested in recognising spontaneous conversation between people, anticipating future multimodal systems that communicate with humans in a human-like manner, rather than more structured dialogue.

3. THE GAZE-CONTINGENT ASR SYSTEM

3.1. Formalising recognition

For speech and visual attention, we can express the decoding problem using probability calculus and Bayesian inference:

$$\begin{aligned} p(v,w|e,y) &\propto p(v,w)p(e|y,v,w)p(y|v,w) \\ &\propto p(v,w)p(y|e,v,w)p(e|v,w) \end{aligned} \quad (1)$$

Where e and y are the set of modality measurements $m = \{e, y\}$, representing the sequence of feature vectors for eye

movement and speech respectively. v and w are the sequence of class types that represent the visual attention sequence and the word sequence respectively.

We can simplify expression 1 by making the following conditional independence assumptions:

- $p(y|e, v, w) = p(y|v, w)$ and $p(e|y, v, w) = p(e|v, w)$: The effect of one modality on another is only via the other's classification, and not the measurement of the other modality itself.
- $p(y|v, w) = p(y|w)$ and $p(e|v, w) = p(e|v)$: The effect of visual attention and speech on one another is via the joint probability $p(w, v)$.
- $p(v, w) = p(w|v)p(v)$: The temporal precedence and the relative confidence in the modality decoding schemes dictates the direction of the dependence; this decides whether to maximise visual attention or word sequence first during decoding.

These assumptions yield $p(w, v)$ in terms of the class conditional probability of a word given the focus of visual attention, $p(w|v)$, and the prior for visual attention, $p(v)$. Thus, for maximum likelihood:

$$p(\widehat{v}, \widehat{w}|e, y) \propto \max_{w} p(w|\widehat{v})p(y|w)p(\widehat{v})p(e|\widehat{v}) \quad (2)$$

Where:

$$\widehat{v} = \arg\max_{v} p(v)p(e|v) \quad (3)$$

3.2. Baseline ASR

To uncover the word sequence $\widehat{w}$ in expression 2, a typical Hidden Markov Model (HMM) large vocabulary continuous ASR system was built. The system was trained using the WSJCAM0 [4], BNC [5], and HCRC Map Task [6] corpora. The system was benchmarked against the standard WSJCAM0 5k test sets and showed credible performance of 20.8% WER.

3.3. Visual-attention classifier

Uncovering the visual attention $\widehat{v}$ in expression 2 from the eye tracking data is straightforward; fixation events, recorded by the eye tracker are assigned to the nearest potential focus of visual attention:

$$\sigma_t = \arg\min_{\varsigma} D(v_t, v^{\varsigma}) \quad (4)$$

Where σ_t is the closest landmark to the visual attention that temporally corresponded to the word onset at time t, and $D(v_t, v^{\varsigma})$ is the Euclidean distance between the landmark ς and the visual attention position v_t at time t.

3.4. Integrating gaze

To make the baseline ASR system gaze-contingent, the probability $p(w|\widehat{v})$ in expression 2 is estimated by modifying the language model probability distribution $p(W)$ at time t to a visual attention-specific language model $p_{\sigma_t}(W)$.

The baseline bigram language model $p(W)$ is constructed using frequentist estimates of bigrams based on their occurrence in the BNC and HCRC map task corpora. Back-off weights are used for the robust estimation of unseen bigrams. The modification of the baseline language model to create an attention-specific language model probability distribution, $p_{\sigma_t}(W)$, is realised by shifting probability mass away from unigrams and bigrams that do not involve words associated with the current focus of visual attention, toward unigrams and bigrams that do.

Let W^{σ} be the set of keywords associated with the visual focus σ, and $W^{\bar{\sigma}}$ be the set of all other words in the language model. Let m be the proportion of mass shifted from each word. New unigram probabilities, $P_{\sigma_t}(W_i)$, were calculated from $P(W_i)$:

$$P_{\sigma_t}(W_i) = \begin{cases} (1-m)P(W_i) & \text{if } W_i \in W^{\bar{\sigma}} \\ P(W_i) + \dfrac{mP(W_i)}{\sum_{W \in W^{\sigma}} P(W)} \sum_{W \in W^{\bar{\sigma}}} P(W) & \text{if } W_i \in W^{\sigma} \end{cases} \quad (5)$$

A similar expression may be obtained for bigram probabilities.

4. EVALUATION

4.1. Matched eye movement and speech data

To evaluate the systems, a set of eye movement direction data and related spontaneous speech was collected for a human-to-human dialogue. The candidate task that participants undertook was loosely based on the HCRC Map Task corpus [6]. There are two participants; an 'Instruction Giver' who describes a geographical map comprising a number of landmarks and a route around them, and an 'Instruction Follower' who recreates the map. Neither participant can see the other, communicating via microphone and headphones only. The Instruction Giver's visual attention was measured using a head-mounted eye tracker. Figure 1 shows an example map superimposed with the frequency distribution of eye of movements; the darkened areas of the map around the landmarks and route indicate high concentrations of the Instruction Giver's visual attention. Nine participants took part in the experiment; all were British nationals and spoke English as their first language. In total 18 sessions were recorded. Session durations ranged from 5-15 minutes.

A commercial head-mounted binocular eye tracker 'EyeLink', was used for tracking the instruction givers' eye position in relation to the map image on a computer monitor.

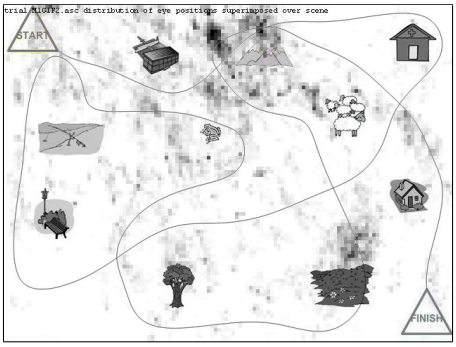

Fig. 1. A frequency histogram showing the distribution of the subject's eye movements over the map.

G: top left triangle with the word start in it
F: ok
G: right now bottom left triangle with the word finish in it
F: yes
G: right if you go from the start in the middle of the page should be a herd of four or five sheep

Fig. 2. Sample transcript from the matched eye and speech data. 'G' indicates the Instruction Giver's speech, 'F' the Instruction Follower's.

Participants' voices were recorded on separate audio channels. The audio and eye movement data captures were synchronised by tagging the eye movement data with audio sample counts during recording. Two passes of the data were made by different human transcribers. Because the human transcribers did not encode all pauses between words, time-aligned transcriptions were regenerated by forced alignment; using the baseline ASR with a language model corresponding to the transcribed word sequence for each speech segment. The speech collected was spontaneous and informal, with disfluencies present. Speakers would stutter, talk over one another, and speak quickly. Figure 2 shows an example of a typical dialogue.

Seven sessions were used for evaluating the gaze contingent ASR; providing 1330 segments of Instruction Giver's speech. We rejected eye data due to non-linear horizontal and/or vertical offsets in gaze direction and losses in the gaze signal identified from the data. For further details of this dataset, see [7].

4.2. Language model type

ASR systems have difficulty in recognising common, shorter length words used in conversational speech (e.g. 'it', 'if' and 'ok'). Shifting probability mass towards visual attention-related words from all other words in the vocabulary may improve the recognition of visual attention-related words but at the expense of all other words in the vocabulary, potentially leading to a rise in overall WER; to avoid this, we optionally shifted only probability mass between words associated with the visual field, leaving the other words untouched. To validate this approach two types of language model are implemented - the 'vocabulary' and the 'visual-field' models. The 'vocabulary' model shifts probability mass towards visual attention-related words from all other words in the vocabulary. The 'visual-field' model shifts probability mass to the visual-attention related words from the set of words associated with the overall visual field. Thus, we hoped to see whether using a gaze-contingent ASR system utilising the latter would be superior. 99% of the probability mass was shifted; this value was determined empirically.

We realised two gaze-contingent ASR systems: System A used the 'vocabulary' language model type and System B used the 'visual field' language model type.

4.3. Tests performed

The baseline ASR system generated a 250-best list for each speech segment. For each gaze-contingent ASR system, this list was reordered after rescoring each word sequence according to the sequence of visual-attention specific language models; identified from the subjects' visual attention at the time of each word onset. Two standard measures of performance were used: Word Error Rate (WER) and the Figure Of Merit (FOM). FOM measures keyword spotting accuracy averaged over 1 to 10 false alarms per hour per keyword; it is a useful complement to the more frequently used WER because it bases its estimate of performance on the ability to detect words pertinent to the communicative intent, regardless of their frequency of occurrence in common language. The resulting measures (based on the most probable word sequence in each list after re-ordering) for the gaze-contingent ASR systems were compared to those of the baseline system.

5. RESULTS

The overall question in this study is whether recognition robustness can be improved by integrating gaze into an ASR system. As the results reveal in table 1, the answer depends on the language model type. The baseline ASR system (column 2) achieved a WER (row 2) of 52.3%. Compared with the baseline, system A (column 3) showed a statistically significant increase in overall WER (row 3); $+3.1$ (p=0.00, n=1330); using the 'vocabulary' language model had the undesirable effect of reducing the recognition accuracy for disfluency-prone words anticipated in section 4.2. System B (column 4) showed increases in WER that were not statistically significant; $+0.5$ (p=0.05, n=1330); in using the 'visual field' language model the recognition of disfluencies-prone words was not compromised.

Performance	Baseline ASR	Gaze-contingent ASR A	Gaze-contingent ASR B
WER(%)	52.3	55.4	52.8
p	-	0.00	0.05
FOM (%)	57.5	55.1	60.3
p	-	0.32	0.40
TP (%)	74.2	79.8	69.4
p	-	0.00	0.02
FA (%)	12.2	17.0	9.1
p	-	0.01	0.05

Table 1. Performance for the ASR systems on collected speech segments (n=1355). The 2-tailed t-test for paired samples was used to test significance.

System B also demonstrated an improved FOM (row 3); $+2.8\%$ ($p = 0.40, n = 42$). In contrast, for system A the FOM decreased; -2.4% ($p = 0.32$). Compared to WER, the FOM results have a lower statistical significance (row 4) and the baseline ASR system recognised the majority of keywords without having to use the information from gaze.

Because the improvements in FOM had low statistical significance, we looked at the component measures of FOM, the keyword spotting True Positives (TP) and False Alarm (FA) counts (rows 5 to 8). The changes in these measures were statistically significant and examining them enabled us to ask whether the desirable increase in FOM for system B was due to an increase in TP and a reduction in FA, as one would desire in a gaze-contingent ASR. The results do not show this: both the TP and FA counts fell for system B; the increase in FOM was due to a fall in both the TP and FA rate compared to the baseline; -4.8% (p=0.02) and -3.1% (p=0.05) respectively.

The observed fall in TP for system B lead to the key finding that integrating gaze did not, on average, lead to improvements in the recognition of words associated with the focus of visual attention. What the integration did was to lower the word probabilities for keywords associated with the visual field except those associated with the focus of visual attention, causing the reduction in both the TP and FA. By definition the number of words associated with the visual field is greater than the number associated with any one visual focus, thus the reduction in the FA count is what makes the gaze-contingent ASR useful.

6. CONCLUSION

Although the baseline ASR system performed well enough to recognise occurrence of the majority of keywords without having to use information from gaze, we have shown that conversational speech recognition can be improved by probabilistically reducing the recognition of words associated with the visual field except those associated with the current focus of visual attention; we recommend this approach for designing gaze-contingent ASR. For evaluating the performance of speech-centric multimodal systems, we have demonstrated the utility of keyword spotting metrics where the keywords represent the communicative intent.

We constrained our evaluation to varying the language model type; choosing the correct language model type is essential for designing successful systems. A direct performance comparison with previous gaze-contingent ASR systems is of limited value; shared datasets may benefit future studies.

Various extensions could be made. The attention-driven language models shifted a fixed proportion of the language model's probability mass to words associated with a specific visual focus from other words. Refinements to this approach would involve learning the temporal asynchrony between modalities and the amount of probability mass to shift. Contaminating the acoustic signal with noise would give more scope for performance improvements. Tracking the eye movement against a dynamic visual field is an extension towards practical realisation which would require robust scene understanding.

7. REFERENCES

[1] A. Hyrskykari, P. Majaranta, and K.J. Räihä, "From gaze control to attentive interfaces," in *Proc. 11th Int. Conf. Human-Computer Interaction (HCII2005)*. 2005, IOS Press.

[2] A. Potamianos, E. Fosler-Lussier, E. Ammicht, and M. Perakakis, "Information seeking spoken dialogue systems part II: multimodal dialogue," *IEEE Trans. Multimedia*, vol. 9, no. 3, pp. 550–566, 2007.

[3] 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.

[4] J. Fransen, D. Pye, T. Robinson, P. Woodland, and S. Young, "WSJCAM0 corpus and recording description," *Cambridge University Engineering Department (CUED) Speech Group, Tech. Rep., September*, 1994.

[5] L. Burnard, "Users reference guide for the British National Corpus," Tech. Rep., Technical report, Oxford University Computing Services, 2000.

[6] A. H. Anderson et al., "The HCRC map task corpus," *Language and Speech*, vol. 34, no. 4, pp. 351–366, 1991.

[7] N. Cooke and M. Russell, "Using the focus of visual attention to improve automatic speech recognition," in *INTERSPEECH '2005 - 9th European Conference on Speech Communication and Technology, Lisbon, Portugal*, 9 2005, pp. 1213–1216.