# SALIENCE BASED LEXICAL FEATURES FOR EMOTION RECOGNITION

*Kalani Wataraka Gamage*[1,2], *Vidhyasaharan Sethu*[1], *Eliathamby Ambikairajah*[1,2]

[1] The School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney NSW 2052, Australia

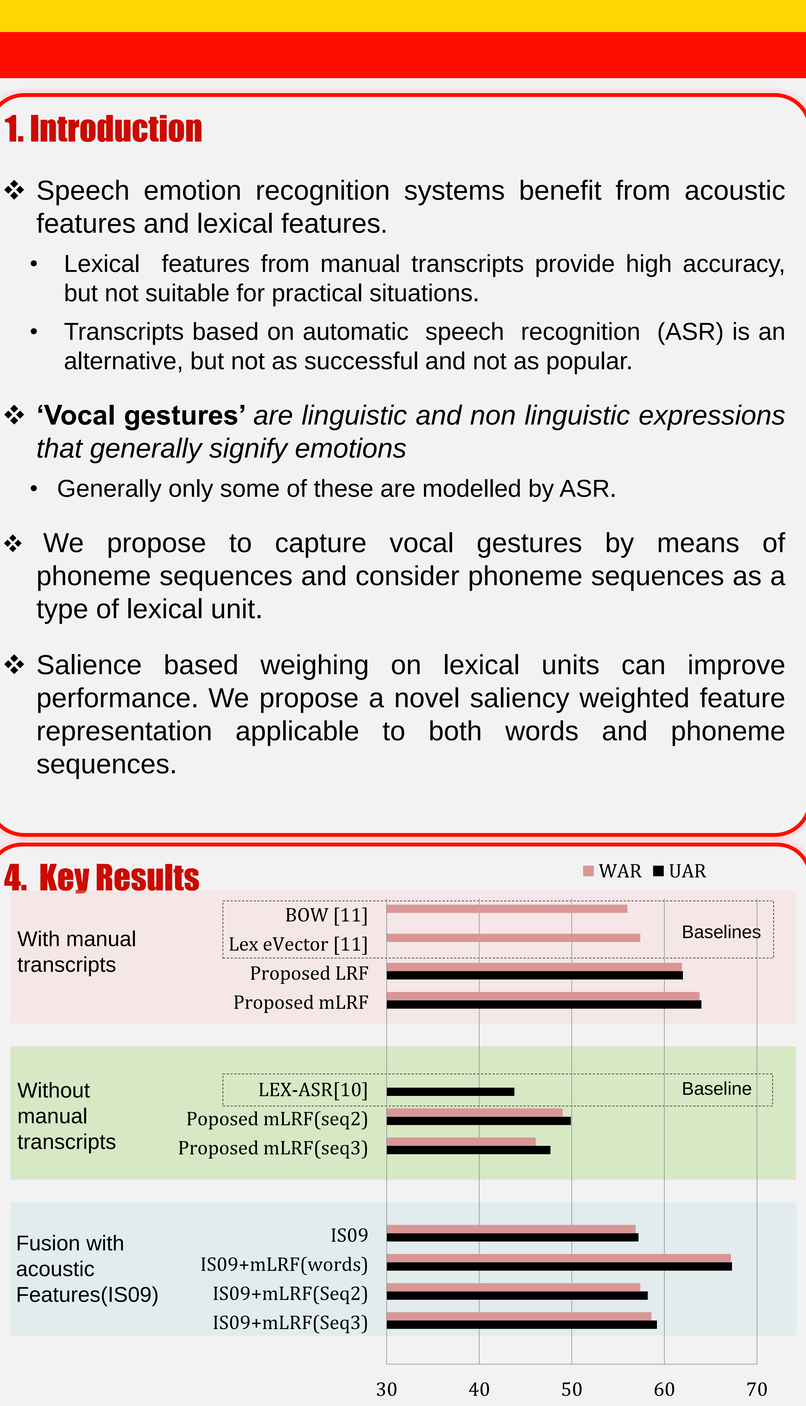
[2] ATP Research Laboratory, Data61, CSIRO, Australia, Sydney NSW 2015, Australia

kalani.watarakagamage@unsw.edu.au

## 1. Introduction

- Speech emotion recognition systems benefit from acoustic features and lexical features.
  - Lexical features from manual transcripts provide high accuracy, but not suitable for practical situations.
  - Transcripts based on automatic speech recognition (ASR) is an alternative, but not as successful and not as popular.
- **'Vocal gestures'** *are linguistic and non linguistic expressions that generally signify emotions*
  - Generally only some of these are modelled by ASR.
- We propose to capture vocal gestures by means of phoneme sequences and consider phoneme sequences as a type of lexical unit.
- Salience based weighing on lexical units can improve performance. We propose a novel saliency weighted feature representation applicable to both words and phoneme sequences.

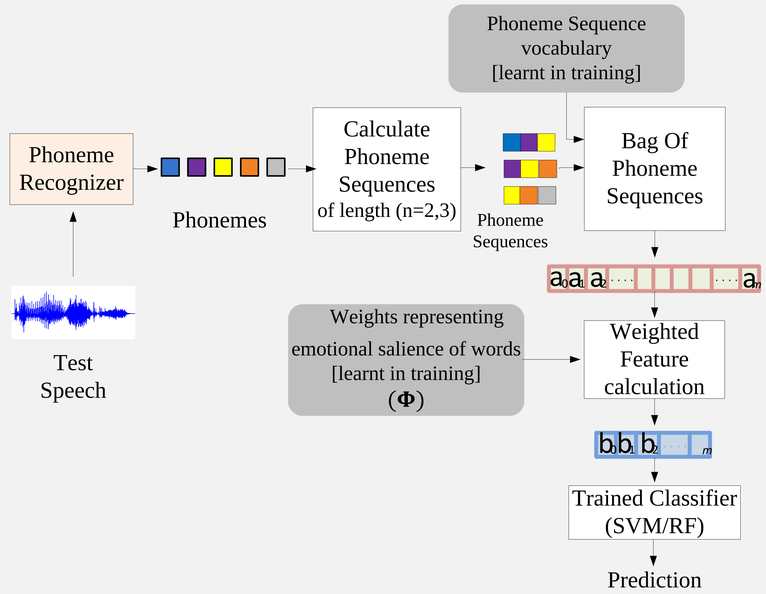
## 4. Key Results




Experiments: **IEMOCAP data on leave-one speaker-out cross validation**

*[Results reported for Proposed features is classified with Random Forest classifier. Lex-ASR [10]-non linear SVM classifier & BOW[11],Lex eVector[11]-linear SVM]*

## 2. Proposed Phoneme Sequence Based Features




- We propose the use of phoneme sequences to encode specific sounds to represent verbal gestures and some spoken content.

## 5. Examples of most salient phoneme sequences

| | |
|---|---|
| Angry | 'aa-hh-hh', 'sh-iy-iy', 'eh-aa-hh', 'iy-iy-n', 'iy-jh-iy', 'ay-ay-t' |
| Happy | 'ow-hh-iy', 'ay-ay-hh', 'hh-aw-aa', 'iy-ae-ay', 'hh-ay-eh' |
| Sad | 'm-m-m', 'n-n-m', 'pau-m-m', 'm-r-hh', 'pau-n-n', 'm-t-ah' |

***examples:***

- '*eh-aa-hh'* (expression of disgust)
- '*iy-ae-ay*' (cheering)
- '*m-m-m'* (sound of long audible breaths or sighs)

## 3. Saliency weights and representations

**Salience based weighing, $v_e$ for utterance $u$**

$$v_e(u) = \frac{1}{K}\mathbf{\Phi} b_p(u)$$

$\mathbf{\Phi}$ is the saliency weight matrix corresponding to each word and each emotion for all words in Bag of Word / Phoneme vector $b_p(u)$ and K is the number of words in utterance $u$.

**The proposed relative frequency based weighing (LRF)**

$$\phi_{j,k} = \frac{\eta_j(w_k)}{1 + \frac{\hat{\eta}_j(w_k)}{n-1}}$$

$\eta_j(w_k)$ : *number occurrences of the $k^{th}$ lexical unit $w_k$, in utterance $u$ of emotion $j$*

$\hat{\eta}_j(w_k)$ : *number occurrences of the $k^{th}$ lexical unit $w_k$, in utterance $u$ of emotion $j$*

$n$ : *the total number of emotions of interest*

**Novel Saliency based lexical feature vectors**

- **LRF : $v_e(u)$ feature based on above weight ($n \times 1$)**
- **mLRF : $\overline{v_e}(u)$ Modified LRF features ($2n \times 1$)**

$$\overline{v_e}(u) = [v_e(u)^T \;\; m_e(u)^T]^T$$

$$where \;\; m_e(u) = [\, a_1(u), a_2(u), .., a_n(u)] \; and \; a_j(u) = \max_i \phi_{j,i}$$

- Unlike Lex eVector, Proposed LRF weight does not apply penalty to words infrequently appearing on different emotion classes.
- Inclusion of $m_e(u)$ in mLRF avoids watering down of the weights in long utterances with a small number of emotionally salient words.

## 6. Conclusion

- Proposed saliency based lexical representation outperforms state-of-the-art lexical features for emotion recognition.
- Phoneme Sequence based features can capture vocal gestures in emotion recognition systems and this approach does not require a full ASR