Enabling Private and Non-Intrusive Smartphone Calls with LipTalk

Muyuan Li, Si Chen, Kui Ren
Department of Computer Science and Engineering
University at Buffalo
{muyuanli, schen23, kuiren}@buffalo.edu

***Abstract*—The typical usage for phones is making calls. However, under certain scenarios, phone calls are inappropriate or intrusive when people are having a meeting, impractical when background noise level is too high or insecure under monitoring of other personals/parties. Typical solution to these problems is to send text-based messages. Yet, we argue that the most natural, efficient way for human-beings to communicate is via speech. In this work, we provide envision for our project, a non-intrusive, convenient and secure communication system. We utilize the front camera of smart phones, efficiency of cloud computing environment and state-of-the-art text-to-speech methods to create a robust visual speech recognition system to enable people to chat with lip movements. We present the current challenges, system architecture, initial findings and planned approaches to our problem.**

***Index Terms*—Non-intrusive communication, Visual speech recognition**

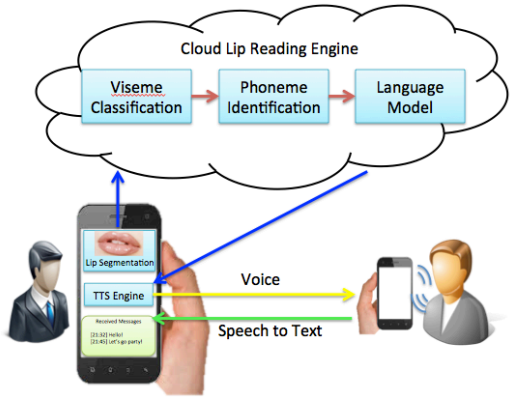


Fig. 1: System Overview

I. Introduction

Since the first day of invention, a speaker and a microphone are the most fundamental components of a phone. For more than a century in the past and infinite many years in the future, phones served, is serving and will serve as a tool for chatting. With the advancement of technology, exciting gadgets are now added to most of the smart phones including cameras, breathtakingly sharp and colorful LCD screens *etc*. At this era, various ways are available for communication. Apart from phone calls, people are able to send text messages, compose multimedia mails or make video chats.

However, there are a wide variety of scenarios in which voice based chatting is not feasible. Typical scenarios include presence in a conference where making phone calls seems rather impolite, inappropriate or even intrusive, in crowded spaces like subway where background noise makes voice impossible to hear and that callers are concerned with phone calls being overheard by other personals or parties. Under most of these circumstances, the best practice nowadays is to send text-based messages. Yet, we argue that, the most natural and efficient way for human-beings to communicate is via speech.

In our project, we aim to deliver a natural, efficient, robust and non-intrusive communication system for smart phones. Instead of the traditional practice of talking via microphone, we explore the possibility of doing real-time visual speech recognition (lip reading) with front cameras that can be found almost on any smart phone. As shown in Figure 1, in our system, people "speak" to front camera. Their lip movements are captured and streamed to cloud for real time identification. The TTS engine translates identification results to sound and deliver it to the other side. Upon receiving voice, the speech identification engine automatically turns voice into text that is displayed on the screen.

Being a research subject for 30 years, a large number of works are devoted to lip reading techniques [1], [2], [3], [4], [5], some of which achieve excellent identification accuracy. However, most of these works only focus on small vocabulary set (e.g. 10 phrases) instead of a practical infinite vocabulary chatting system. To enable such lip reading, we are exposed to various challenges, the most important of which are:

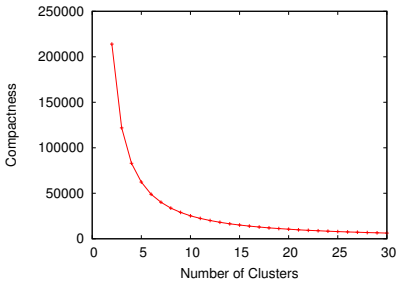
1) We focus on facilitating a new app that will utilize advanced ML techniques and practical observations to help people communication in real time without disturbing others. Hence the user experience and accuracy are the major challenges;
2) Human voice is produced with a combination of vocal cords, tongues, resonance cavity, teeth and lip. These information are available in MRI-based visual speech recognition [6], but in our setting, only lip movements and teeth are visible. In large vocabulary set, such confusion tend to have greater impact.

We differentiate our project from previous works that traditional lip reading targets at different application scenarios that is not adequate. In our study, traditional lip reading technique is a basic building block, but we greatly improve it by making new observations in our application scenarios:

1) Most people tend to speak slower or make more visible lip movement when speak without producing voice;
2) We reproduce speech via TTS engines and deliver it via mobile network. Hence, some confusion in lip identification such as *short* and *shot* is acceptable. We believe human are able to address such ambiguity given proper context

In addition, we adopt a cross-domain approach that integrates solutions from multiple different disciplines are adopted to improve the overall usability and accuracy.

II. Our Approach on Identification



(a) Picking K in K-Means Clustering

(b) Centroid of Each Cluster

Fig. 2: Viseme Clustering

The identification engine contains a pipeline of machine learning algorithms. Vision algorithms are adopted to identify face and mouth regions, after which we classify visemes and reproduce phoneme sequence, synthesizing with language models to achieve better result.

Fig. 3: Sequence of Visemes Forming Phonemes

A. Mouth Region Segmentation

Our mouth region segmentation uses a chain of state-of-the-art Haar-like Cacade Classifiers [7] already implemented in OpenCV. We perform face detection and mouth region identification in the lower 1/3 region of face.

B. Viseme Classification

There are a variety of phoneme to viseme mappings in previous works [8]. Instead of grouping phonemes into visemes, We assign each static frame with a viseme label. We run K-Means algorithm [9] over videos of 30 min obtained from VOA Learning English. Figure 2a indicates that the optimal number of clusters is either 7 or 8. Figure 2b shows the centroid of each cluster. In this way, phonemes are represented by a series of visemes (see Figure 3, the vowel AW is represented by a sequence of viseme label in Figure 2b). Unlike previous works that concatenate fixed numbers of frames [3] for viseme classification, our approach only studies one picture at a time and naturally captures temporal information by assigning different viseme series to phonemes. More precisely, the viseme series can be viewed as observations and a Hidden Markov Model [10] (HMM) is trained to classify viseme sequences into phonemes.

C. Additional Language Model Layer

Interestingly, the HMM model is one of the best practices in speech recognition works [11], [12]. Hence, we incorporate one additional HMM layer trained from plain text materials to address ambiguity of phonemes and compensate for possible missing parts during prediction.

III. Conclusion

There are many scenarios in which we need to chat without voice. Text messages are just inadequate in that they are less efficient and unnatural as compared to speech. Unfortunately, current visual speech recognition techniques do not fit the needs. Hence, in this paper, we give envision of our project on practical lip chatting application. We have presented its use scenarios, potential challenges and our technical approach of the identification algorithms. We expect this project to be eventually delivered as an intuitive easy-to-use service that makes non-intrusive, efficient, natural phone calls a reality.

References

[1] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. S. Huang, "Avicar: audio-visual speech corpus in a car environment." in *INTERSPEECH'04*, 2004.
[2] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *Multimedia, IEEE Transactions on*, vol. 11, no. 7, pp. 1254–1265, 2009.
[3] T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass, "A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments," in *ICMI'04*, 2004.
[4] Z. Zhou, G. Zhao, and M. Pietikainen, "Towards a practical lipreading system," in *CVPR'11*, 2011.
[5] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for hmm based automatic lipreading," in *ICIP'98*, 1998.
[6] S. Narayanan, E. Bresch, P. K. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. C. Lammert, M. I. Proctor, V. Ramanarayanan, and Y. Zhu, "A multimodal real-time mri articulatory corpus for speech research." in *INTERSPEECH'11*, 2011.
[7] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR'2001*, 2001.
[8] S. Hilder, B.-J. Theobald, and R. Harvey, "In pursuit of visemes," in *AVSP'10*, 2010.
[9] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. California, USA, 1967.
[10] L. Rabiner and B. Juang, "An introduction to hidden markov models," *ASSP Magazine, IEEE*, vol. 3, no. 1, pp. 4–16, 1986.
[11] K.-F. Lee and H.-W. Hon, "Large-vocabulary speaker-independent continuous speech recognition using hmm," in *ICASSP'88*. IEEE, 1988.
[12] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.