INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

An Investigation of Crowd Speech for Room Occupancy Estimation

Siyuan Chen[1], Julien Epps[1,2], Eliathamby Ambikairajah[1,2], Phu Ngoc Le[1,2]

[1]School of Electrical Engineering and Telecommunications
The University of New South Wales, UNSW Sydney, NSW 2052, Australia
[2]Data61 CSIRO, Australia

siyuan.chen@unsw.edu.au,j.epps@unsw.edu.au,e.ambikairajah@unsw.edu.au,phule@unsw.edu.au

Abstract

Room occupancy estimation technology has been shown to reduce building energy cost significantly. However speech-based occupancy estimation has not been well explored. In this paper, we investigate energy mode and babble speaker count methods for estimating both small and large crowds in a party-mode room setting. We also examine how distance between speakers and microphone affects their estimation accuracies. Then we propose a novel entropy-based method, which is invariant to different speakers and their different positions in a room. Evaluations on synthetic crowd speech generated using the TIMIT corpus show that acoustic volume features are less affected by distance, and our proposed method outperforms existing methods across a range of different conditions.

Index Terms: Occupancy estimation, crowd speech, babble noise, speaker count, entropy, acoustic volume

1. Introduction

Using speech techniques for room occupancy estimation is a relatively new research area in the speech community, even though the more general problem of occupancy estimation has already gained attention in the building construction and management area [1,2]. The aim of occupancy estimation is to count and track people in a room for HVAC (i.e., heating, ventilation and air conditioning) control systems, whose settings can be adjusted automatically accordingly. Case studies report that such demand-driven HVAC control can reduce commercial office building energy costs by about 70% [1], making sustainable smart buildings possible in the future.

Current proposed occupancy estimation techniques involve a variety of sensors, including CO_2 concentration, temperature, humidity, ultrasonic, image, light, sound, electromagnetic signals, power meter, computer app and chair sensors [1,2]. Each can provide an extent of comprehensive and fine-grained occupancy information, as surveyed in [1]. However, none of them can work alone to provide the full detail of occupancy information for every measurement circumstance.

Audio and speech based systems for occupancy estimation have the advantages of non-intrusiveness, real-time estimation, cost-effectiveness, and ability to protect user privacy, which make them promising to investigate. Currently, there are few publications [2] on this research, although large projects have been established to employ acoustic-based systems in FPGA platforms for occupancy estimation [3]. They are focused on hardware implementation, while the efficacy of existing speech-based methods for room occupancy estimation is unknown.

In this paper, we specify the occupancy estimation problem as accurately counting the number of people speaking simultaneously in a room without tracking. To our best knowledge, this is the first effort to investigate occupancy estimation in terms of small/large crowds and distance sensitivity under party-like conditions, e.g. in a party or a classroom, using speech-based methods. We assume that there is only one microphone placed in a room for a low implementation cost, and no personal identity detection is required. We investigate the influence of distance change, since speaker locations in a room in reality may change from one use context to the next. We also propose new distance-independent features which can improve estimation accuracies.

2. Background and Related Work

An ideal occupation estimation system should be able to count the number of people speaking reliably, regardless of what is being said, who is speaking or where they are in a room, as opposed to (for example) speaker diarization [10,11] which determines who spoke when. This is a non-trivial problem because the changes of speakers, linguistic information and speaking locations in a room can cause large variations of acoustic features used to model speech.

An important assumption for this party-mode speech-based occupancy estimation [2] is that all people are speaking simultaneously, so that the estimated speaker count is the room occupancy. However, this assumption is not realistic. A manual analysis conducted in [4] showed that the distributions of number of speakers speaking per frame were different in two scenarios: two conversations consisting of four speakers versus two individual speakers speaking simultaneously. They found that the most probable number of speakers speaking at the same time during the two conversations was two. Therefore, speech-based techniques can only estimate the number of people speaking. Only when every speaker was participating continuously in conversation was the estimated number close to the true room occupancy.

The first study to explore the feasibility and performance of speech-based occupancy estimation in smart building environments was probably [2]. Their approach took the room shape (round and rectangular), microphone location (central and corner), crowd speech duration (5, 10, 15, 20, 25 s) and crowd size (5, 10, 20, 40, 80 speakers) into account, and employed a background Short-Term Energy (STE) technique to estimate room occupancy. Specifically, a STE histogram was constructed for the speech utterance of each crowd size. A nonparametric kernel-smoothing curve fitting was applied to each histogram to obtain its STE mode. Estimation accuracy was then evaluated by estimating the overlapped area between

http://dx.doi.org/10.21437/Interspeech.2017-70

the Gaussian distributions of STE mode for different speaker count. Although this method achieved nearly 100% accuracy for five classes of crowd speech, there are two major concerns. Firstly, the performance was not evaluated with unseen data. Secondly, STE might be sensitive to the distance between speakers and the microphone.

Another relevant study [4] comes from quite a different application – an investigation of babble speech noise modeling and analysis. Their aim was to improve in-set/out-of-set speaker verification by choosing the correct babble noise model. Hence, the room setting for crowd speech was not considered. The authors only tested their approach for small crowds of 1 to 9 speakers, so the applicability to large room occupancy estimation is unknown.

This babble speaker count method seems attractive for occupancy estimation because the authors demonstrated that the acoustic volume enclosed by their GMMs (Gaussian Mixture Models) centroids was reduced as the number of speakers increased in babble. This phenomenon is caused by less distinct phones being presented when there are more speakers speaking at the same time. Mel-frequency cepstral coefficients (MFCC) were used to characterize the acoustic space by clustering GMMs in each frame. Then DCT (Discrete Cosine Transform) vectors of the probability distribution of each speaker count were trained to estimate the final babble speaker count.

We hypothesize that, unlike STE mode, the babble speaker count method will be less affected by the distance between speakers and the microphone.

3. Methods

3.1. Speaker count estimation based on STE mode

In the STE mode method, which we treat as a baseline method, the energy of the *i*th frame of speech x_i is firstly calculated as [2]

$$E_i = \frac{1}{N}\sum_{n=1}^{N} w(n)|x_i(n)|^2 \quad (1)$$

where w is a window function of length N samples. Then the distribution of energy $f(E)$ for a crowd speech recording (S) is obtained by a nonparametric kernel-smoothing curve fitting to the histogram of E. STE mode value is estimated as [2]:

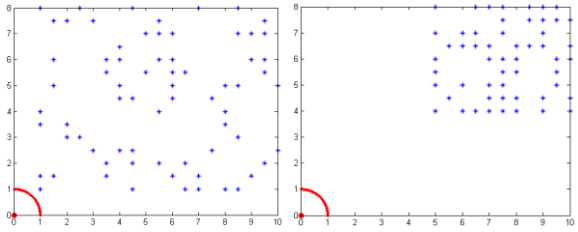
$$STE_{mode} = arg\max_E f(E) \quad (2)$$

Different to [2], where the statistical model is not learned and tested, a speech-based occupancy estimation system which can learn from data is preferred. Therefore we train models (M) for C classes using STE_{mode} normalized by the max value, and then classify speaker count C_s according to maximum-likelihood estimation, using STE_{mode} normalized by the max value in the training data.

$$C_s = arg\max_c p(S|M_c) \quad (3)$$

3.2. Speaker count estimation based on babble noise modeling

In the first stage of the babble speaker count estimation framework described in [4], each frame of a crowd recording S was classified into C classes using maximum-likelihood, based on one GMM model per class adapted from a universal background model (UBM) with d-dimensional MFCCs, whose mixtures characterize the acoustic volume. Then an N-bin



(a) (b)

Figure 1. *Simulated room settings for speech-based occupancy estimation, with a left corner microphone denoted by the red solid dot. Each speaker is at least 0.5 m away from each other. There is no speaker within one meter to the microphone, indicated by the red curve. (a) An example of 80 speakers positioned uniformly in the room (close). (b) An example of 80 speakers positioned uniformly in the back corner (remote) to test how the distance to the microphone affects each method.*

histogram h of speaker count C for the crowd recording S was constructed, and a k-dimensional DCT was applied to the histogram. The DCT coefficients exhibit reduced correlations relative to the histogram, and can be used to reduce the dimension of the representation.

$$D(k) = \sqrt{\frac{2}{N}}\sum_{n=1}^{N} h(n)\cos\left(\frac{\pi}{2N}(2n-1)(k-1)\right) \quad (4)$$

where $k = 2, 3, \ldots, N$.

In the second stage, new models for the C classes were represented by the DCT feature vectors. The final speaker count C_s' was determined by the correlation of the test DCT feature vector ($\boldsymbol{D}$) and the DCT feature vector ($\boldsymbol{D}_c$) calculated from the training data for speaker count C [4].

$$C_s' = \text{argmax}_c \frac{cov(\boldsymbol{D},\boldsymbol{D}_c)}{\sigma_D \sigma_{D_c}} \quad (5)$$

3.3. Proposed speaker count estimation based on entropy

Intuitively, when more people are speaking at the same time, the chance of recognizing a distinct single phoneme becomes lower. Therefore, a straightforward way is to observe the probability change of each phoneme using a phoneme detector. However, just like the findings shown in [4] that MFCC can capture the information of acoustic volume and acoustic movement change, we found that our proposed entropy based method can capture the information of phoneme probability change without a phoneme detector. We hypothesized that for any given frame of speech, if it comprises only a single speaker, the GMM posterior probability should tend to be large for only a few mixture components, and small for others. By contrast, if multiple speakers are present, the posterior probability values should tend to be more even across all the mixture components. Thus the room occupancy count could be indicated by a measure of the "spread" of posterior probabilities, which we capture using entropy, and the variance of entropy across frames.

To have a prior, we construct a UBM [8] from crowd speech recordings with different speaker counts. Given a d-dimensional MFCC feature from the ith frame (x_i), the probability associated with the jth GMM-UBM mixture (w_j, μ_j, Σ_j) is p_j.

$$p_j = w_j \frac{1}{\sqrt{|\Sigma_j|(2\pi)^d}} e^{-\frac{1}{2}(x_i-\mu_j)'\Sigma_j^{-1}(x_i-\mu_j)} \quad (6)$$

After the probabilities are normalized, the entropy of the ith frame (H_i) is

$$H_i = -\sum_{j=1}^{J} p_j log_2 p_j \quad (7)$$

where $\sum p_j = 1$ and J is the number of mixtures. Two features were developed for each crowd speech: the mean entropies (H_{mean}) and the standard deviation of entropies (H_{std}) across frames. Maximum-likelihood estimation was also used to classify speaker count.

4. Experimental Work

4.1. Database

We employed the TIMIT corpus [5] for this investigation for comparability with [2], and used the designated training and testing speakers to train and test models for each method. The training and testing data contained disjoint speakers and comprised 462 and 168 speakers respectively, where each speaker recording was constructed from 10 TIMIT utterances randomly ordered to form 25s of speech, since this duration has the best accuracy in [2]. The background noise in audio files was removed by examining the energy before processing. We used the training data to create 600 synthetic crowd speech recordings for each class as the training data and another 600 to train a UBM, and generated 300 crowd speech recordings using the testing data for each class as the test data.

Similarly to [2], crowd speech (S) of speaker count (C) was artificially generated with decay modeling using I individual TIMIT recordings (x), with a distance of r_i between the ith speaker and the microphone, and amplitude A_i. Thus, distance was taken into account as:

$$S = \sum_{i=1}^{I} A_i x_i \quad (8)$$

$$A_i = \frac{r_0}{r_i} A_0 \quad (9)$$

where r_0 was set to 1 meter and A_0 was set as 1. r_i is a uniformly and randomly generated value between 1 meter and the size of the room in Figure 1.

4.2. Experimental configurations

For the purpose of classification comparison, we employed *small crowd* classes of {1, 2, 3, 4, 5, 6, 7, 8, 9} speakers as in [4] and *large crowd* classes of {5, 10, 20, 40, 80} speakers as in [2], respectively. The rectangular room setting (8 × 10m) in [2] was used, as shown in Fig. 1(a). However we limited the distance between two speakers to be no less than 0.5m [6] in order to make it more realistic, rather than assuming that average speaker density is no more than one speaker per square meter [2].

To estimate room occupancy count, we (i) trained models with the two entropy features, H_{mean} and H_{std}, as described in Section 3.3, and performed classification as shown in equation (3); and (ii) concatenated the feature vectors of H_{mean} and H_{std} to the DCT feature vector D_c in the babble speaker count approach to improve the estimation accuracy.

4.3. Evaluation on small and large crowds

Apart from estimating the accuracies for large crowd in [2], we also tested STE_{mode} for a small crowd. Similarly, we evaluated the babble speaker count approach not only for small speaker count [4] but also for large speaker count. All three methods from Section 3 employed a Hamming window, and other parameter settings were the same as [2] and [4], i.e., 50ms frame size in the STE mode method [2] and 125ms frame size, 19-dimensional MFCC (cepstral mean normalized [13] in our study), 32 mixture GMMs and 10-dimensional DCT in the babble speaker count method [4]. The number of bins in the histogram is 24 in this study, including both the small and large crowd classes. For the maximum-likelihood GMM classifier [7], the number of mixtures was 1 for STE mode, and 3 for the entropy-based method.

4.4. Distance sensitivity investigation

To address the distance sensitivity question, we used the room setting shown in Fig. 1(b) to create new crowd speech recordings as the test data, and examined how estimation accuracies change due to this influencing factor.

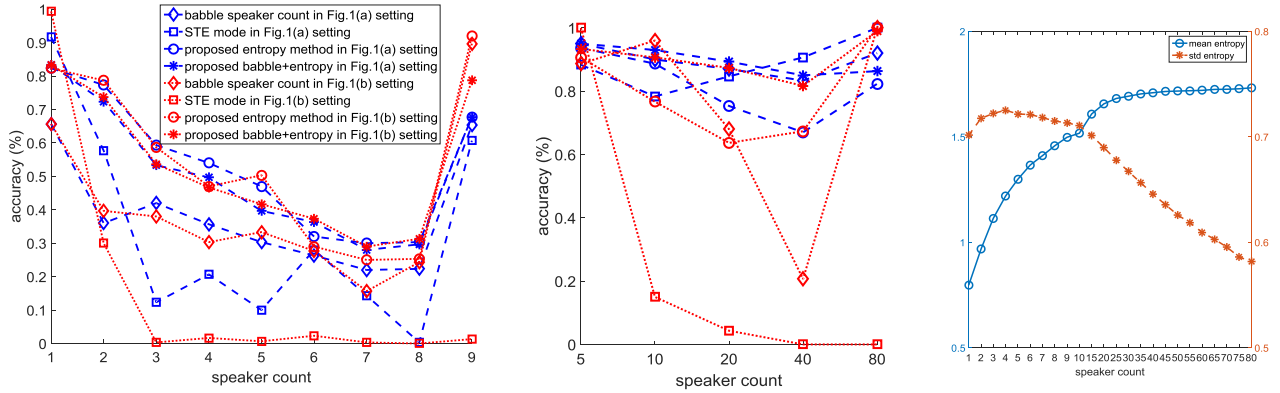
5. Results and Discussion

To verify our hypotheses, the performance of the STE mode and babble speaker count methods, which are the baselines for this study, together with our proposed entropy based methods were evaluated under the same four conditions: small and large crowd size in close and remote-microphone distance. Fig. 2(a) shows the accuracies of the four methods for *small* crowd size classification in the two room settings shown in Fig. 1. Among them, the method denoted as babble + entropy method is to examine the complementarity of the two methods by concatenating entropy features into DCT feature vector in the second stage of the babble speaker count. Fig. 2(b) presents the accuracies of the same four methods for *large* crowd estimation in the same two room settings.

Table 1: *Overall speaker count estimation accuracies (%) for different microphone distance and crowd size settings.*

Method	Close (Fig. 1(a))		Remote (Fig. 1(b))	
	9-class (small)	5-class (large)	9-class (small)	5-class (large)
STE mode [2]	32.8	88.4	15.1	23.9
Babble speaker [4]	38.4	89.5	40.5	74.7
Entropy (proposed)	**53.3**	81.4	**54.3**	79.7
Babble + Entropy	51.1	**89.7**	52.8	**90.4**

It can be observed that the proposed entropy method and Babble + Entropy provide strong classification accuracy across all conditions, although the entropy method is slightly weaker than others for the large crowd close-microphone problem. By contrast, the STE mode method seemed to have reasonable accuracy only for large crowd and not fine-grained occupancy estimation. This is probably because with a large increment in speaker count, the amount of energy change is as noticeable as that of acoustic space features for large crowd. Meanwhile, the two-stage babble speaker count method (new in this context) is able to estimate both small and large crowd size but its small crowd estimation performance is ordinary as shown in Table 1. However, the fusion of entropy features and babble DCT features in the second stage of the babble speaker count method improves the estimation performance compared to the babble speaker count method only, as shown in Table 1.



(a) (b) (c)

Figure 2. *Classification accuracy for (a) small crowd and (b) large crowd in two room settings, where "+" denotes fusion, 'close' refers to Fig. 1(a) setting and 'remote' refers to Fig. 1(b) setting. The horizontal axis tells the number of classes. (c) shows the average value of entropy features versus 24 speaker count classes for speaker count of 1 to 9 and 10 to 80 with an increment of 5.*

Together with the two entropy based methods showing the best performance under most conditions, it suggests that the proposed entropy features are able to distinguish different speaker count regardless of the size of the crowds.

In terms of distance sensitivity, as indicated by Fig. 2(a) and (b), it is as expected that the estimation performance of the STE mode method dropped significantly when speakers moved to the back corner of the room, whereas the other methods were more tolerant of the distance change. This is not surprising, because when speakers are far away from the microphone, they have less contribution to the energy, but they may have similar contributions to acoustic volume and the phoneme overlaps, which show up in the babble and entropy features. From this investigation, we can observe that our proposed entropy based method provides the best accuracies among previous studies and is currently the most robust method under the variations of the number of speakers and their locations in a room.

Fig. 2(c) provides a big-picture understanding of why the 2-dimensional entropy feature vector can achieve reasonably good estimation accuracy. We can see that the mean entropy, H_{mean}, increases monotonically as the number of speaker increases from 1 to 80. As explained in Section 3.3, the entropy feature is an alternative to measure phoneme overlap using a phoneme detector. Our additional experiment using a phoneme detector [14,15] confirmed that the entropy of the obtained phoneme posterior probabilities has the similar results to Fig. 2(c). With so many speakers, the acoustic space is cluttered and the frame-level [9,12] entropy is always high across many frames. The more speakers contribute their speech simultaneously in a crowd speech, the more frames have large entropy value.

Furthermore, we can observe that the mean entropy, H_{mean}, saturates when the number of speakers increases beyond around 30, which means that the feature becomes less discriminative for large crowds. Interestingly, the variation of entropy across frames provides another picture. Fig. 2(c) demonstrates that the standard deviation of entropy, H_{std}, increases as the speaker count increases slightly and then begins to decrease. For a single speaker, the variation among neighboring frames is small. When there are more speakers, this variation increases. For larger speaker counts, we speculate that the frame-level entropy is always high, and therefore the standard deviation of frame-level entropy is small.

It is worth noting that the analyses of small and large crowd sizes employed different numbers of trained classes. Their chance level accuracies are different (11% and 20% respectively), therefore, we cannot directly compare the accuracy of speaker count of 5 in both the small and large crowd size analysis. Meanwhile, the edge classes such as 1 and 9 in small crowd size analysis and 5 and 80 in large crowd size often achieved very high accuracies. This is very likely because these classes have only one neighbouring class where errors mostly occur, while other classes often have two, observed from their confusion matrices. Furthermore, the synthetic crowd speech data did not fully represent real-world conditions, since for example room reverberation and different spoken languages were not considered. These are worth further investigation in the future.

6. Conclusions

In this paper, we investigated two current methods and proposed a third one for room occupancy estimation in two key problem configurations of interest: small and large crowds, and close and remote microphone placements. From the perspective of an ideal speech-based occupancy estimation system, which is invariant to crowd size and room microphone positioning, we found that our proposed entropy based method outperformed other methods in previous studies. This seems to be because these two entropy features successfully represent acoustic space characteristics that are less influenced by the size of the crowd and their locations. Future work involves augmentations of speech-based occupancy estimation in more real-world environments, such as rooms with reverbaration issues and people speaking in different languages.

7. References

[1] T. Labeodan, W. Zeiler, G. Boxem, and Y. Zhao, "Occupancy Measurement in Commercial Office Buildings for Demand-driven Control Applications – A Survey and Detection System Evaluation", *Energy and Buildings*, pp. 303-314, 2015.

[2] Q. Huang, Z. Ge, and C. Lu, “Occupancy Estimation in Smart Buildings Using Audio-Processing Techniques”, *ICCCBE*, 2016.
[3] B. Kelly, D. Hollosi, P. Cousin, S. Leal, B. Iglar, and A. Cavallaro, “Application of Acoustic Sensing Technology for Improving Building Energy Efficiency”, *Procedia Computer Science*, pp. 661-664, 2014.
[4] N. Krishnamurthy, and J. H. L. Hansen, “Babble Noise: Modeling, Analysis, and Applications”, *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 1394-1407, 2009.
[5] C. Lopes, and P. Fernando, “Phone Recognition on the TIMIT database”, *Speech Technologies*, pp. 285-302, 2011.
[6] S. Heshka and Y. Nelson, “Interpersonal Speaking Distance as a Function of Age, Sex, and Relationship”, *Sociometry*, pp. 491-498, 1972.
[7] D. A. Reynolds, and R. C. Rose, “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 72-83, 1995.
[8] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models”, *Digital Signal Processing*, pp. 19-41, 2000.
[9] O. Plchot, M. Diez, M. Soufifar, and L. Burget, “PLLR Features in Language Recognition Systems for RATS”, *INTERSPEECH*, 2014.
[10] D. A. Reynolds, and P. Torres-Carrasquillo, “Approaches and Applications of Audio Diarization”, *ICASSP*, 2005.
[11] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, “Speaker Dirarization: A Review of Recent Research”, *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 356-370, 2012.
[12] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, G. Bordel, “On the Use of Phone Log-likelihood Ratios as Features in Spoken Language Recognition”, *IEEE Spoken Language Technology Workshop*, 2012.
[13] A. Rosenberg, C. H. Lee, F. Soong, “Cepstral Channel Normalization Techniques for HMM-based Speaker Verification”, *Proceedings of the International Conference of Spoken Language Processing*, Vol. 4, pp. 1835–1838, 1994.
[14] S. Irtza, V. Sethu, P.N. Le, E. Ambikairajah, H. Li, "Phonemes Frequency Based PLLR Dimensionality Reduction for Language Recognition." *INTERSPEECH*, 2015.
[15] P. Schwarz, “Phoneme Recognition Based on Long Temporal Context,” Ph.D. dissertation, Faculty of Information Technology, Brno University of Technology, http://www.fit.vutbr.cz/, Brno, Czech Republic, 2008.