



HHS Public Access

Author manuscript

Proc IEEE Int Conf Acoust Speech Signal Process. Author manuscript; available in PMC
2023 April 14.

Published in final edited form as:

Proc IEEE Int Conf Acoust Speech Signal Process. 2022 May ; 2022: 6462–6466. doi:10.1109/

icassp43923.2022.9747006

TOWARDS INTERPRETABILITY OF SPEECH PAUSE IN DEMENTIA DETECTION USING ADVERSARIAL LEARNING

Youxiang Zhu¹, Bang Tran¹, Xiaohui Liang¹, John A. Batsis², Robert M. Roth³

¹Department of Computer Science, University of Massachusetts Boston, MA, USA

²School of Medicine, University of North Carolina, Chapel Hill, NC, USA

³Geisel School of Medicine at Dartmouth, Lebanon, NH, USA

Abstract

Speech pause is an effective biomarker in dementia detection. Recent deep learning models have exploited speech pauses to achieve highly accurate dementia detection, but have not exploited the interpretability of speech pauses, i.e., what and how positions and lengths of speech pauses affect the result of dementia detection. In this paper, we will study the positions and lengths of dementia-sensitive pauses using adversarial learning approaches. Specifically, we first utilize an adversarial attack approach by adding the perturbation to the speech pauses of the testing samples, aiming to reduce the confidence levels of the detection model. Then, we apply an adversarial training approach to evaluate the impact of the perturbation in training samples on the detection model. We examine the interpretability from the perspectives of model accuracy, pause context, and pause length. We found that some pauses are more sensitive to dementia than other pauses from the model’s perspective, e.g., speech pauses near to the verb “is”. Increasing lengths of sensitive pauses or adding sensitive pauses leads the model inference to Alzheimer’s Disease (AD), while decreasing the lengths of sensitive pauses or deleting sensitive pauses leads to non-AD.

Keywords

Speech pauses; interpretability; acoustic feature; dementia detection; spontaneous speech

1. INTRODUCTION

Exploiting spontaneous speech for dementia detection has shown promise in recent years. In the Alzheimer’s Dementia Recognition through Spontaneous Speech (ADReSS) challenge [1, 2], spontaneous speech datasets were collected using a Cookie Theft picture description task [3], and researchers have developed deep learning models for achieving a promising classification accuracy in differentiating AD from non-AD patients [4, 5]. One common acoustic biomarker these successful models have exploited is the speech pauses [4, 5, 6], which have been widely studied in the medical research [7, 8]. However, these deep learning models have not analyzed the impact of the positions and lengths of the speech pauses on the detection of dementia. Thus, it remains unknown that to what extent these models understand the speech pauses to make the decision.

Recent deep transfer learning studies have significantly advanced our understanding of the domains of image, language, and speech. Deep transfer learning models learn features automatically from large-scale datasets, and these auto-learned features can be more representative than conventional handcrafted features. In dementia detection, such models produced a high accuracy [9, 4, 5], but they often lacked inter-pretability, limiting the trustworthiness of their medical applications. Attention and visualization mechanisms are common methods to enhance the interpretability of deep learning models [10, 11, 12, 13, 14, 15]. However, these methods do not apply to the speech-based dementia detection scenario because of the complex relationship between input data and labels. In the image domain, labels are directly related to the input data, e.g., a dog and a dog image. Humans can evaluate whether the interpretability results are consistent with human knowledge, e.g., the model assigns large weights to the dog's body area to recognize the dog. In contrast, dementia labels are derived from a standard and complex assessment, which is not directly related to the speech data. Humans hardly evaluate the interpretability results on dementia-related speech. Therefore, we evaluated the interpretability of speech pauses, rather than words expressed, for dementia detection.

The WavBERT dementia detection model [16] is the first model preserving the pause in the text without using the manual transcription. WavBERT uses Wav2vec [17] to measure the lengths of speech pauses and represents them with punctuation marks. In this paper, we use adversarial learning and WavBERT to study the speech pauses. Specifically, we first propose an adversarial attack approach where we generate adversarial samples from original testing samples by adding, replacing, and deleting the punctuation marks. We test the WavBERT model using the adversarial samples and identify the perturbation action that leads to the most significant change in the model's confidence. In an adversarial training approach, other than training with the original samples, we additionally train the WavBERT model with two groups of adversarial samples, one includes samples positively influencing the confidence, and the other includes samples negatively influencing the confidence. We then investigate how the adversarial training affects the model on the testing samples. The contributions of this paper are three-fold.

First, we exploit adversarial learning to study the inter-pretability of speech pauses in dementia detection. We propose methods to evaluate the impact of speech pauses from the perspectives of model accuracy, pause context and length.

Second, we found that the speech pauses between "is" and another verb is dementia-sensitive, and confirmed this observation with the training set. We also found that adding pauses at sensitive places or increasing the lengths of sensitive pauses leads the model inference toward AD, while deleting sensitive pauses or decreasing the lengths of sensitive pauses leads the model inference toward non-AD.

Third, we found that the misspelling errors from Wav2vec introduced uncommon tokens, leading to misinterpretation of the speech pauses. Such problem is inevitable if using a speech recognition model and a language model, and might be resolved with large-scale speech representation.

2. WAVBERT - PAUSE PRESERVATION

WavBERT [16] utilizes speech pauses and transcripts for dementia detection. It transcribes the speech audio to transcript using Wav2vec, and this transcript does not have punctuation marks. WavBERT uses the number of blank tokens from the intermediate results of Wav2vec to define two levels of pauses: longer sentence-level pause and shorter insentence pause. A “period” mark is inserted into the transcript if a sentence-level pause is determined; a “comma” mark is inserted into the transcript if an in-sentence pause is determined. After adding the punctuation “period” and “comma” into the transcripts, WavBERT feeds the transcripts to BERT for AD classification. More details can be found in Figure 1 and the paper [16]. In this paper, we add one more level of in-sentence pause and represent it with “semicolon.” The length of such pause is longer than “comma” and shorter than “period.”

3. PROPOSED ADVERSARIAL LEARNING

We present two adversarial learning approaches to exploit the impact of speech pauses on dementia detection.

Adversarial attack.

An adversarial attack adds perturbation to the punctuation marks. Note that, other adversarial attacks on BERT considered the entire vocabulary as the attack space [18, 19]. In our case, the attack space only includes punctuation marks, and the perturbation actions are adding, deleting, and replacing. As shown in Figure 1(b), we adopt the same training phase of the original WavBERT. To launch the attack, we choose an original testing sample s_0 , and generate a set of adversarial samples by using a single action of adding, deleting, or replacing. For adding, we add one punctuation mark between two neighboring words where no punctuation mark exists; for deleting, we delete an existing punctuation mark; and for replacing, we replace an existing punctuation mark with a different punctuation mark, e.g., replacing “comma” with “period.” We define a **confidence level** as $l_{AD} = o_{AD}(s)$ for an AD sample s , and $l_{non-AD} = o_{non-AD}(s)$ for a non-AD sample s , where $o_{AD}(s)$ and $o_{non-AD}(s)$ are the logit output by the model for AD and non-AD labels. Among the samples generated from s_0 , we identify a sample s_1 as the most effective sample if the confidence level of s_1 is the lowest. We call s_1 as the step-1 sample, and then iteratively launch the attack on s_1 to obtain the step-2 sample s_2 . The attack will be launched for a maximum of 20 steps unless the confidence level cannot be lowered. The details are shown in Algorithm 1.

Algorithm 1 Adversarial attack $s_i \rightarrow s_{i+1}$

Input: A step i sample s_i and a model M
Output: A step $(i + 1)$ sample s_{i+1}

- 1: $s_{i+1} = \text{NULL}$, confidence level $l_i = M(s_i)$
- 2: Generate a set of adversarial samples from s_i using one perturbation action: $C_{i+1} = \{s_{i+1,1}, s_{i+1,2}, \dots, s_{i+1,n}\}$.
- 3: **for** each sample $s_{i+1,j}$ in C_{i+1} **do**
- 4: $l_{i+1,j} = M(s_{i+1,j})$.
- 5: **if** $l_{i+1,j} < l_i$ **then**
- 6: $l_i = l_{i+1,j}$, $s_{i+1} := s_{i+1,j}$
- 7: **Return** s_{i+1}

Adversarial training.

In another adversarial learning approach, we generate adversarial samples and apply adversarial training to the WavBERT, as shown in Figure 1(c). Specifically, we add perturbations to the original training samples of WavBERT. The perturbation actions are the same as the previous adversarial attack approach. In each attack step on a sample, we choose the adversarial sample that has the **lowest** confidence level. The attack is iteratively launched for a maximum of 20 steps unless the confidence level cannot be lowered. Finally, the original training samples and perturbed samples are used to train the WavBERT. The updated model is expected to be less effective because adversarial samples produce a negative impact in the training phase. We further propose a reversed adversarial training approach. We generate an adversarial sample from an original sample that has the **highest** confidence level. In this way, we envision that the reversed adversarial training makes the model more effective.

Interpretability.

We study the interpretability of speech pauses from the following three perspectives.

Model accuracy.—In the adversarial attack approach, we can evaluate how much the perturbation affects the model accuracy. In the adversarial training approach, the changes of model accuracy reveal how well the model understands the speech pauses in the original samples.

Pause context.—In each step of the adversarial attack, we choose the perturbation that causes the most negative impact on the confidence level. We can analyze the statistics of the neighboring words of all the chosen punctuation marks, and identify the dementia-sensitive pauses based on the context.

Pause length.—We divide the perturbation actions into two groups. Group 1 involves adding and replacing a short pause with a long pause, and Group 2 involves deleting and

replacing a long pause with a short pause. Group 1 shifts the result towards AD, and Group 2 shifts the result towards non-AD.

4. EVALUATION

In this section, we present the implementation details and report our evaluation results on speech pauses.

4.1. Implementation details

We followed the original implementation of WavBERT [16] using PyTorch. The training and the adversarial training used the standard training and testing sets of ADReSS speech only (ADReSSo). The training was conducted with 10 rounds, and the average result of 10 rounds was reported. Compared to the original setting, the learning rate was increased to 10^{-5} , the maximum epoch was reduced to 100 for faster convergence, and the model accuracy remains similar [16].

4.2. Model accuracy of adversarial attack/training

Figure 2 shows the impact on accuracy from adversarial attack and adversarial training. The original WavBERT used a majority vote of 10 rounds and achieved an accuracy of 83-84%. The average accuracy of 10 rounds was 82-83%. We chose to report the average accuracy because it reflects the performance of every individual round. As shown in Figure 2(a), with the step-1 attack, the accuracy is lowered to 70-72%, which shows the WavBERT model may easily be impacted by small perturbation. However, humans in their speeches often add pauses for many unknown reasons, which are not related to their cognitive problems. In addition, we observed that if a multi-step attack adds more perturbation on pauses, the accuracy is significantly lowered. We confirmed that the pause is an important factor for model inference. Lastly, the level-3 model resulted in the lowest accuracy when the step-20 attack was launched. We consider that the level-3 model relies more on pauses to make the inference, and thus is more vulnerable to the adversarial attacks on pauses. Figure 2(b) and (c) show the accuracy of models using adversarial training and reversed adversarial training was lowered but not significantly, thus revealing WavBERT's limited understanding of pauses.

4.3. Context analysis of pauses

Table 1 shows the context analysis results of perturbation. We listed the top five contexts based on the frequency for range (1, 2) and attack step (1, 5, 20). The adversarial samples listed under the column step- k include all the adversarial samples generated from step-1 to step- k . We have three observations: i) We found the pause between “is” and a verb is highly dementia-sensitive, as shown in the highlighted green. We confirmed that the long pause between “is” and a verb only appears in the AD samples but not non-AD samples of the training set. ii) We found interviewers' pause is important, as shown in the underlined text. As this type of pause does not appear in the step-1 attack, it is not considered as the most important pauses for model inference. We consider it still affects the accuracy to some extent. Perez-Toro et al. has reported a similar result [20]. iii) We found ASR misspelling errors are important, such as “othing tote # ofor” and “gon as # aa itre”, as shown in the

oval boxes. We consider the misspelling errors introduce the uncommon tokens, which are considered as important context in the downstream task because the pretrained models have limited knowledge about these tokens.

4.4. Length analysis of pauses

Table 2 shows the frequency of perturbation actions. Adding and deleting the long pauses (punctuation “period”) are common perturbation methods to change detection results (highlighted). In step-1 attack, we consider the perturbation actions are performed on sensitive pauses. In cases of changing confidence towards non-AD, the total number of deleting and long-to-short replacing actions is significantly larger than the total number of adding and short-to-long replacing. In cases of changing confidence towards AD, the perturbation actions are mostly adding pauses. However, we note that most actions of adding a comma in level 2/3 and adding a period in level 1 lead to a confidence change towards non-AD. Thus, we consider that not all pauses are dementia-sensitive from the model’s perspective, because natural language contains pauses. Lastly, we found that replacing actions appear less frequently than adding and deleting, suggesting that the pause context is more important than the pause length.

5. CONCLUSIONS

In this paper, we exploited adversarial learning to study the interpretability of speech pauses in dementia detection. With the adversarial attack and training techniques, we successfully identified the positions and lengths of sensitive pauses, e.g., speech pauses near to the verb “is” are dementia-sensitive. We also confirmed that increasing lengths of sensitive pauses or adding sensitive pauses leads the model inference to AD, while decreasing the lengths of sensitive pauses or deleting sensitive pauses leads to non-AD. Our evaluation reveals that the misspelling errors are inevitable if using speech recognition and language model. One future direction is to exploit large-scale speech data for pre-training a speech-based language model, thus resolving the misspelling problem.

Acknowledgments

This research is funded by the US National Institutes of Health National Institute on Aging, under grant No. R01AG067416.

LIST OF ACRONYMS

AD	Alzheimer’s Disease
ADReSS	Alzheimer’s Dementia Recognition through Spontaneous Speech
ADReSSo	ADReSS speech only

REFERENCES

- [1]. Luz Saturnino, Haider Fasih, de la Fuente Sofia, Fromm Davida, and MacWhinney Brian, “Alzheimer’s dementia recognition through spontaneous speech: the adress challenge,” arXiv preprint arXiv:2004.06833, 2020.

- [2]. Luz Saturnino, Haider Fasih, de la Fuente Sofia, Fromm Davida, and MacWhinney Brian, “Detecting cognitive decline using speech only: The addresso challenge,” arXiv preprint arXiv:2104.09356, 2021.
- [3]. Goodglass Harold, Kaplan Edith, and Weintraub Sandra, BDAE: The Boston Diagnostic Aphasia Examination, Lippincott Williams & Wilkins Philadelphia, PA, 2001.
- [4]. Yuan Jiahong, Bian Yuchen, Cai Xingyu, Huang Jiaji, Ye Zheng, and Church Kenneth, “Disfluencies and fine-tuning pre-trained language models for detection of alzheimer’s disease.,” in INTERSPEECH, 2020, pp. 2162–2166.
- [5]. Yuan Jiahong, Cai Xingyu, Bian Yuchen, Ye Zheng, and Church Kenneth, “Pauses for detection of alzheimer’s disease,” *Frontiers in Computer Science*, vol. 2, pp. 57, 2020.
- [6]. Sluis Rachel A, Angus Daniel, Wiles Janet, Back Andrew, Gibson Tingting, Liddle Jacki, Worthy Peter, Copland David, and Angwin Anthony J, “An automated approach to examining pausing in the speech of people with dementia,” *American Journal of Alzheimer’s Disease & Other Dementias*[®], vol. 35, pp. 1533317520939773, 2020.
- [7]. Pastoriza-Dominguez Patricia, Torre Ivan G, Dieguez-Vide Faustino, Gomez-Ruiz Isabel, Gelado Sandra, Bello-Lopez Joan, Avila-Rivera Asuncion, Matias-Guiu Jordi, Pytel Vanesa, and Hernandez-Fernandez Antoni, “Speech pause distribution as an early marker for alzheimer’s disease,” medRxiv, pp. 2020–12, 2021.
- [8]. Davis Boyd H and Maclagan Margaret, “Examining pauses in alzheimer’s discourse,” *American Journal of Alzheimer’s Disease & Other Dementias*[®], vol. 24, no. 2, pp. 141–154, 2009.
- [9]. Balagopalan Aparna, Eyre Benjamin, Rudzicz Frank, and Novikova Jekaterina, “To bert or not to bert: comparing speech and language-based approaches for alzheimer’s disease detection,” arXiv preprint arXiv:2008.01551, 2020.
- [10]. Ribeiro Marco Tulio, Singh Sameer, and Guestrin Carlos, “‘‘ why should i trust you?’’ explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.
- [11]. Goyal Yash, Wu Ziyang, Ernst Jan, Batra Dhruv, Parikh Devi, and Lee Stefan, “Counterfactual visual explanations,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 2376–2384.
- [12]. Simonyan Karen, Vedaldi Andrea, and Zisserman Andrew, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” arXiv preprint arXiv:1312.6034, 2013.
- [13]. Clark Kevin, Khandelwal Urvashi, Levy Omer, and Manning Christopher D, “What does bert look at? an analysis of bert’s attention,” arXiv preprint arXiv:1906.04341, 2019.
- [14]. Liu Nelson F, Gardner Matt, Belinkov Yonatan, Peters Matthew E, and Smith Noah A, “Linguistic knowledge and transferability of contextual representations,” arXiv preprint arXiv:1903.08855, 2019.
- [15]. Che Zhengping, Purushotham Sanjay, Khemani Robinder, and Liu Yan, “Distilling knowledge from deep networks with applications to healthcare domain,” arXiv preprint arXiv:1512.03542, 2015.
- [16]. Zhu Youxiang, Obyat Abdelrahman, Liang Xiaohui, Batsis John A, and Roth Robert M, “Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection,” *Proc. Interspeech 2021*, pp. 3790–3794, 2021.
- [17]. Baevski Alexei, Zhou Henry, Mohamed Abdelrahman, and Auli Michael, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” arXiv preprint arXiv:2006.11477, 2020.
- [18]. Garg Siddhant and Ramakrishnan Goutham, “Bae: Bert-based adversarial examples for text classification,” arXiv preprint arXiv:2004.01970, 2020.
- [19]. Li Linyang, Ma Ruotian, Guo Qipeng, Xue Xiangyang, and Qiu Xipeng, “Bert-attack: Adversarial attack against bert using bert,” arXiv preprint arXiv:2004.09984, 2020.
- [20]. Pérez-Toro PA, Bayerl SP, Arias-Vergara T, Vásquez-Correa JC, Klumpp P, Schuster M, Nöth Elmar, Orozco-Arroyave JR, and Riedhammer K, “Influence of the interviewer on the automatic assessment of alzheimer’s disease in the context of the addresso challenge,” *Proc. Interspeech 2021*, pp. 3785–3789, 2021.

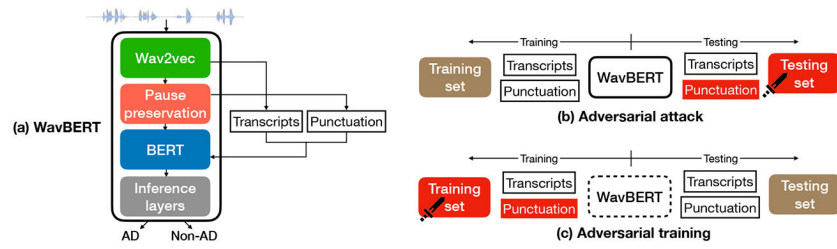


Fig. 1: The description of WavBERT model (a), the adversarial attack approach (b), and the adversarial training approach (c).

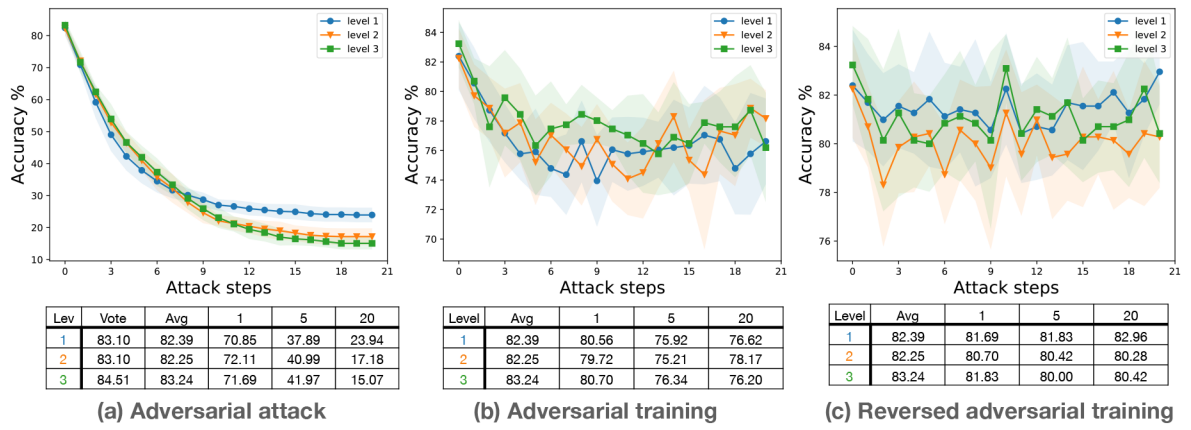


Fig. 2: Impact of adversarial attack/training. Level 1: period. Level 2: comma + period. Level 3: comma + semi-colon + period

Table 1:

Context analysis on frequency of pauses. ‘#’ represents a pause. r represents the range, and s represents the step

	$r = 1, s = 1$	$r = 1, s = 5$	$r = 1, s = 20$	$r = 2, s = 1$	$r = 2, s = 5$	$r = 2, s = 20$
	it # looks	19 of # the	40 the # sink	102 the # sink	10 the mother # and the	17 the picture # tell me
	of # the	16 dishes # and	32 and # the	71 and # the	10 othing tote # ofor	13 mother is # washing dishes
level 1	dishes # and	12 it # looks	27 of # the	60 of # the	10 ah it # looks like	20 ah it # looks like
	mother # and	11 i # see	25 the # little	54 the # little	10 it is # o	12 the picture # tell me
	tote # ofor	10 is # drying	19 i # see	50 i # see	10 gon as # aa iire	18 washing dishes # and the
	of # the	17 is # running	38 and # the	117 and # the	10 or cake # and look	17 is open # looks like
	it # looks	15 dishes # and	38 dishes # and	93 the # sink	10 sink is # running over	37 mother is # washing dishes
level 2	is # drying	13 of # the	37 on # the	92 on # the	10 the mother # and the	18 washing dishes # and the
	mother # and	12 and # the	32 of # the	76 of # the	10 othing tote # ofor	18 the picture # well the
	is # running	11 the # sink	27 i # see	66 i # see	10 ah it # looks like	18 of water # and the
	it # looks	15 dishes # and	40 and # the	120 and # the	10 or cake # and look	18 and the # sink is
	of # the	12 of # the	35 the # sink	96 the # sink	10 ah it # looks like	37 mother is # washing dishes
level 3	cake # and	10 is # going	31 on # the	93 on # the	10 little girl # cubboard doors	19 washing dishes # and the
	dishes # and	10 is # running	30 of # the	78 of # the	9 othing tote # ofor	23 washing dishes # and the
	girl # cubboard	10 and # the	30 dishes # and	78 dishes # and	9 what wit # the sinks	20 water on # the floor
					9 stool is # going to	20 the picture # well the
						20 and the # sink is

Table 2:

Length analysis. A stands for AD, N for Non-AD

	Perturbation	Step-1		Step-5		Step-20	
		A→N	N→A	A→N	N→A	A→N	N→A
level 1	Add.	90	357	495	1764	1336	5715
	Delete .	242	3	844	34	1168	644
level 2	Add,	34	1	272	35	1631	1365
	Add .	38	334	157	1639	450	4410
	Replace ,→.	5	16	58	86	196	502
	Delete ,	57	9	388	19	1501	369
	Delete .	180	0	591	9	879	188
	Replace .→,	36	0	284	12	653	243
	level 3	Add ,	36	2	284	15	1686
Add;		14	21	58	123	303	1009
Add .		34	311	111	1524	329	4118
Replace ,→;		0	5	4	9	89	141
Replace ,→.		2	10	19	89	86	454
Replace ;→.		3	2	29	9	72	88
Delete ,		8	7	136	13	946	218
Delete ;		41	0	187	0	444	75
Delete .		153	2	488	11	719	163
Replace ;→,		9	0	92	1	467	64
Replace .→,		42	0	284	5	630	82
Replace .→;		8	0	58	1	169	73