Language independent on-off voice over IP source model with lognormal transitions

Ahmed D. Shaikh[1], Keith J. Blow[1], Marc A. Eberhard[1], and Scott A. Fowler[2]

[1] School of Engineering and Applied Science, Aston University, Birmingham B4 7ET, UK
[2] Department of Science and Technology, ITN, Linköping University Bredgatan-34, SE-601 74, Norrköping, Sweden
E-mail: k.j.blow"aston.ac.uk

Abstract

The recent explosive growth of voice over IP (VoIP) solutions calls for accurate modelling of VoIP traffic. This paper presents measurements of ON and OFF periods of VoIP activity from a significantly large database of VoIP call recordings consisting of native speakers speaking in some of the world's most widely spoken languages. The impact of the languages and the varying dynamics of caller interaction on the ON and OFF period statistics are assessed. It is observed that speaker interactions dominate over language dependence which makes monologue based data unreliable for traffic modelling. The authors derive a semi-Markov model which accurately reproduces the statistics of composite dialogue measurements.

1 Introduction

The increasing popularity and commercial success of Voice over IP (VoIP) solutions has led to a steady rise in VoIP traffic over the past few years. As major network operators migrate from the traditional circuit switched network to the converged IP network, VoIP traffic is poised to grow even further, constituting a large portion of global IP traffic [1]. This calls for accurate modelling of the statistical properties of VoIP calls with emphasis on efficient bandwidth management. VoIP calls have often and traditionally been modelled using on-off source models where periods of speech activity, talk spurts, are represented as ON periods T_{ON}, separated by silence lengths represented as OFF periods T_{OFF} . The voice packet transmission occurs only for the duration of the ON periods and the traffic sources are turned off during the OFF periods of a VoIP call, thus making the correct identification of OFF periods in a VoIP traffic stream, crucial for efficient traffic transmission.

Several on-off models for VoIP calls have been proposed in the past [2]. The most prominent of these [3-8] assumed an exponential or geometric approximation for the T_{ON} and T_{OFF} periods and their results were based on old and outdated measurements of talk spurts and silence lengths from analogue voice call recordings of limited duration. In addition the calls were based on English speakers only. However recent work [9-12] on packetised voice has shown that the T_{ON} and T_{OFF} periods are not exponentially distributed but are rather heavy tailed. While heavy tailed distributions for the T_{ON} and T_{OFF} periods have been proposed [9-12], the short T_{OFF} periods (< 200 ms), which have been shown to be significant [7], were ignored. Consequently much longer T_{OFF} and T_{ON} periods [12] were observed.

In this paper, we present in Section 2 new and accurate measurements of the T_{ON} and T_{OFF} periods, including the short ones, from a very large multi-lingual database of VoIP calls of sufficiently long duration so as to achieve good quality statistical information. We then investigate the impact of the language and speaker behaviour on the T_{ON} and T_{OFF} periods. Finally we compare our results to previous measurements [8] and the prominent two-way voice model [7] which included the short silences of less than 200ms. We show that conventional Markov models cannot accurately reproduce the silence statistics. We then derive a simple semi-Markov model based on lognormal transitions and show that this can accurately reproduce the measured statistics.

2 Measurement of T_{ON} and T_{OFF} periods

2.1 Voice Call Database

For the measurement of T_{ON} and T_{OFF}, we considered a large database of two-way voice call recordings retrieved from the 'Callfriend' section of TalkBank.org [13]. This database consisted of over 100 h of voice call recordings involving both male and female native speakers of a few of the world's most widely spoken languages including Mandarin Chinese, Spanish, Japanese, German, French and English. Each of the call samples typically lasted from 30 to 50 min and have a sampling frequency of 8 kHz and a constant bit rate of 64 kb/s but are otherwise unprocessed (e.g. no silence suppression). Apart from linguistic differences, the call samples featured varying degrees of random dynamics of conversation and the presence of non-grammatical silences as one would observe in a typical telephone conversation. In addition to the large VoIP call database, a selection of multi-lingual monologue data for the aforementioned languages was also considered for the purpose of analysis of short silence lengths, based on the monologue bursts of a speaker in a VoIP call, as was done in the work of [7] as well. The monologue data, lasting up to 15 min for each

language, was retrieved from the Open Speech Repository [14].

2.2 Silence Detection

The measurement process was carried out by extracting the monaural speech signal *S(t)* from the source file and deriving the smooth acoustic envelope of the speech signal for silence discrimination. Note that for the silence detection analysis we only consider the 'conversation part' of the VoIP call as suggested in [12]. Fig. 1 shows an example of a speech signal extracted from a VoIP call sample in the English language. We see that the talk spurts are separated by silence lengths marked by the T_{ON} and T_{OFF} periods respectively.

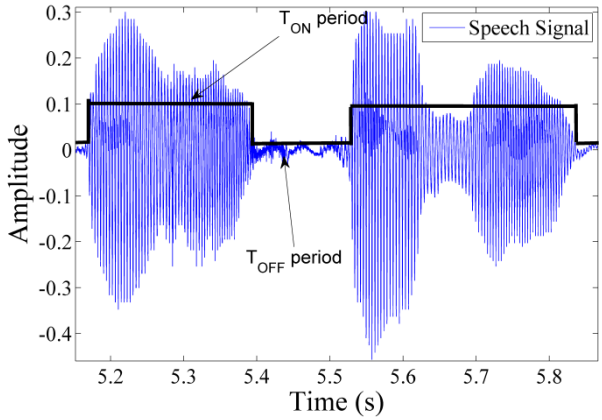

Fig. 1 *Example of speech signal S(t) with T_{ON} and T_{OFF} periods*

We will now describe the method used to identify the T_{ON} and T_{OFF} periods in the speech data. In measuring the statistics of these periods we use continuous time in the sense that no binning into frames is used, this is done later solely for the purpose of comparison with previous studies. The speech signal *S(t)* is first filtered using an efficient FFT low pass filter which has the following frequency response function:

$$H(f)=\begin{Bmatrix}0,|f|>f_c\\1,|f|<f_c\end{Bmatrix},\ f_c=1\,KHz \qquad (1)$$

where f_c is the cutoff frequency of the low pass filter. As the speech signals associated with the VoIP call recordings consisted of very large datasets, we made use of the overlap-add (OLA) method of FFT processing, to avoid memory issues. Since we are only interested in obtaining the envelope of the signal and not reconstructing the speech signal, the choice of f_c was carefully chosen so that all short silence lengths or T_{OFF} periods are tracked as accurately as possible. Fig. 2 shows the cumulative distribution of silence lengths and the impact of different choices for the value of f_c for an arbitrary call sample. We observe that for low values of f_c we see fewer short silences whereas for higher values of f_c we see more short silences as we would expect. For example for a value of f_c = 100 Hz we observe that the results closely match those in [3] where silences below 200 ms were not observed. For a large value of f_c = 2 kHz, we see that noise in the data dominates, thus damaging the statistics. The shortest silence period that can be detected by an adult is approximately 5 ms [15]. Moreover, the choice of standard frame sizes used in the current generation of voice codecs varies from 5 ms to 30 ms [16]. Given this, the value of f_c was chosen so that we record as many T_{OFF} periods as close to 5 ms as possible while limiting the presence of T_{OFF} periods shorter than 1 ms, assuming those to be noise. A cut-off frequency f_c = 1 kHz, as used to generate Fig. 1, proved to be the most appropriate choice for our requirements for the low pass filtering process.

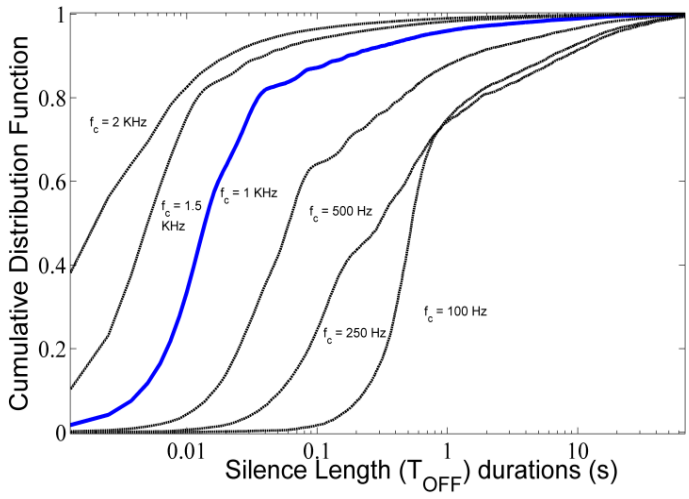

Fig. 2 *Impact of varying f_c on the T_{OFF} periods*

After the low pass filtering, the acoustic envelope *m(t)* is derived by computing the modulus of the filtered signal $S_f(t)$, which is then subjected to a decision making process to discriminate the talk spurts T_{ON} from the silence lengths T_{OFF}, by use of a Threshold Crossing Value (TCV) placed comfortably above the estimated noise floor of the speech signal. The noise floor of course varies with each VoIP call recording sample, but as a general rule of thumb, the TCV is chosen midway between the noise floor and the smallest talk spurt spike that can be detected which stands out from the noise floor, which is approximately 1.5-2.0 % of the maximum amplitude range of the speech signal *S(t)*. This value is then used in a decision making process whereby the duration for which the envelope m(t) lies below the TCV, is marked as a silence length (T_{OFF} period) and the duration for which the envelope m(t) lies above the TCV is marked as a talk spurt (T_{ON} period). Once the values of the T_{ON} and T_{OFF} periods are accumulated for the entire speech signal, we finally compute the probability density function and cumulative density function for statistical analysis. The impact of the choice of the TCV on the statistics of the talk spurts and silence lengths is consistent with previous work [3]. For a lower TCV, we record a larger number of short silences and talk spurts, whereas for a higher TCV, we record fewer and longer silences and talk spurts. Fig. 3 shows the impact of the change in TCV on the cumulative statistics of the T_{OFF} periods. The TCVs are represented as percentages of the maximum positive amplitudes of the speech signal *S(t)*. We can clearly see in Fig. 3 that over a reasonable range of threshold values, the statistics are consistent.

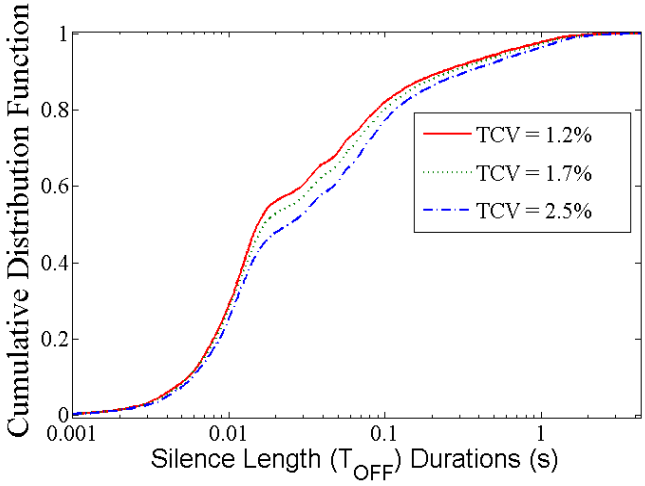

Fig. 3 *Impact of varying TCV on the T_{OFF} periods*

3 Impact of language and prosodic factors

We now consider the impact of the language and the behaviour pattern of the speaker on the statistics of T_{ON} and T_{OFF}.

Since the conventional voice traffic models, including the well-known 8-state model [7], are based on monologue data, we first study the characteristics of the T_{OFF} periods based on the monologue data from [14]. Plotting the average statistics for the monologue samples of each group of languages, it was observed that the incidence of short T_{OFF} periods is unique to each language. An example of this observation is shown in Fig. 4 which shows the probability densities for the silences for three different languages. Whereas we see no major differences in densities for the long and medium silence lengths, we see that Mandarin Chinese has the highest number of short silence lengths (< 200 ms), followed by French and then English. This is because the short silence lengths in spontaneous monologue speech vary with the size of the prosodic units and speech rate involved in the language [17]. Thus we see in our results that Chinese and French, being syllable-timed languages, have a larger number of short silence lengths due to the shorter prosodic units as compared to English, a stress-timed language which has relatively longer prosodic units. This observation is independent of the samples chosen for the language.

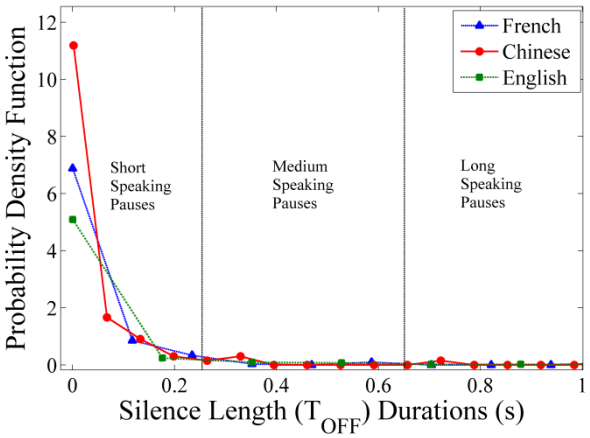

Fig. 4 *Comparison of monologue silences (T_{OFF} periods) for Chinese, French and English Languages*

These differences in the short silence lengths with every language have been proved to be statistically significant [17, 18]. Voice models based on monologue speech should, therefore, take the language into consideration as the speech rate varies with every language. We note that the dependence of the short silence lengths on the language for the monologue samples is independent of the speaker of the same language.

However in our analysis of T_{OFF} periods of dialogue from the large multi-lingual VoIP call database [13], we observe that the statistical differences in languages are overshadowed by the random dynamics of the interaction of the speakers. Fig. 5 shows the cumulative distribution functions (CDFs) for the T_{OFF} periods for calls in different languages. We present the T_{OFF} periods in frames of 5 ms as in [7, 8] to facilitate ease of comparison. We record about 10-30% more short T_{OFF} periods (<100 ms) in our measurements when compared to those in [8].

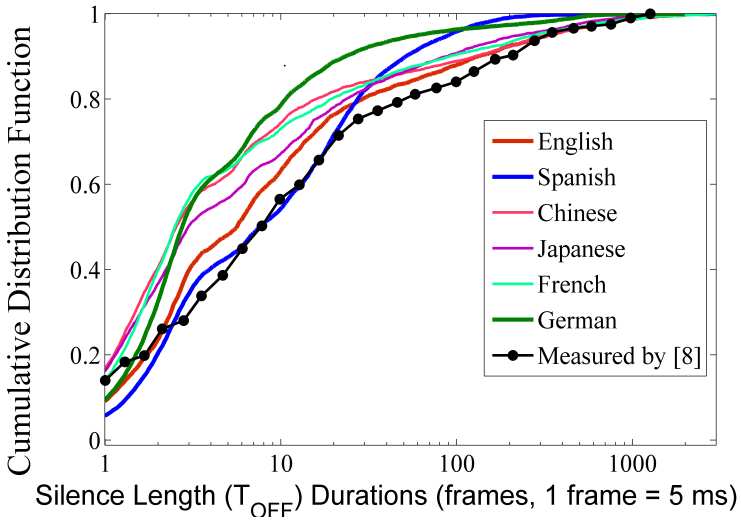

Fig. 5 *Cumulative distribution of T_{OFF} periods of VoIP call samples for various languages*

The results shown in Fig. 5 seem to suggest a significant language dependence of the T_{ON} and T_{OFF} distributions. However, this is not the case as we will now show.

One initial observation that was made from the analysis of the VoIP call database, is that the differences in the cumulative distributions of T_{OFF} periods for call samples of the same language were as large as those of call samples of different languages. This is clearly apparent from the results in Fig. 6 which show the cumulative distributions of T_{OFF} periods for eight different call samples for the Spanish language. As can be seen, the distributions for the various call samples of the Spanish language vary as much as the distributions of the call samples for different languages shown in Fig. 5.

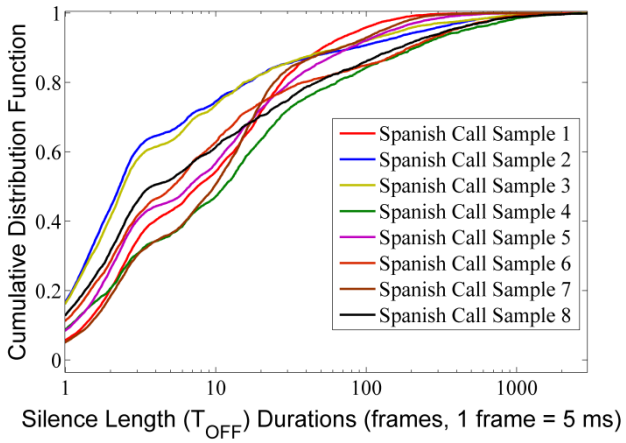

Fig. 6 *Cumulative distribution of T_{OFF} periods of VoIP call samples for the Spanish language*

Another significant observation from the results is that although there are differences in distributions of silence lengths in the voice call samples for each language, the frequent presence of non-grammatical silence lengths in the voice call samples, such as those related to listening, thinking and hesitation, yields a complex anomaly which does not present a viable option to model the voice call silences on a linguistic or a speech rate basis alone. This is clearly observed with the German and Spanish samples which have a large presence of prolonged semi-intentional silence lengths or silences of hesitation, thus yielding higher numbers of medium and long T_{OFF} periods when compared to the other language samples.

A similar observation has been made with the cumulative statistics of the T_{ON} periods. Fig. 7 shows the cumulative statistics for the T_{ON} periods for calls in different languages. The presence of non-grammatical talk spurts makes it almost impossible to characterize the T_{ON} periods on the basis of language. The speech rate with which the caller speaks in a particular language is a more dominant parameter than the prosodic or temporal structure of the language when it comes to the statistical properties of the T_{ON} periods. This speech rate changes from one call to another as a direct result of the varying nature of emotions and psychological behaviour of the speaker. For example, it can be seen that the Spanish call sample has far fewer T_{ON} periods due to the presence of several long silence periods related to hesitation in relation to their talk activity. In addition it is also important to note that the T_{ON} period statistics are also affected by the presence of non-speech expressions such as coughing, laughing, anger and others. Hence there is an absence of uniformity in the observed T_{ON} periods for the various call samples including those of the same language, as was observed in the case of T_{OFF} periods.

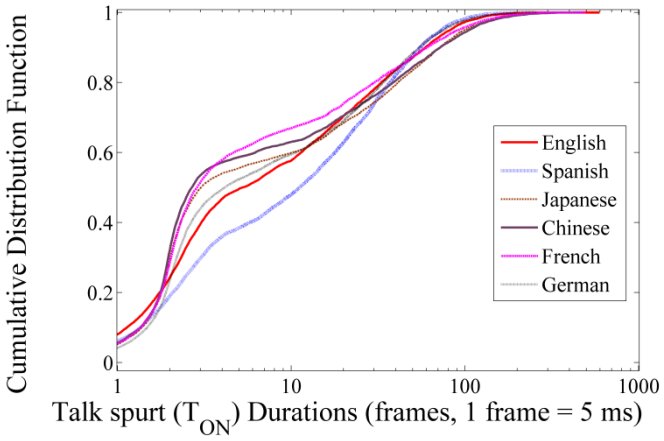

Fig. 7 *Cumulative distribution of T_{ON} periods of VoIP call samples for various languages*

Our study of the impact of language and speaker behaviour on the T_{ON} and T_{OFF} period statistics has revealed that while there are distinguishable differences observed with samples of monologue speech for each language, the presence of random dynamics of conversation and non-grammatical silence lengths and talk spurts in the VoIP call samples yields complex behaviour inconsistent with the modelling of T_{ON} and T_{OFF} periods on a linguistics or a speech rate basis alone, thus leading us to a conclusion: VoIP statistics are unaffected by the languages spoken but rather differ only based on the random dynamics of conversation. In other words, the observed VoIP statistics are independent of the language spoken. Therefore in order to propose an appropriate model which can represent the conversational characteristics and dynamics of all the call samples in the database, it is imperative that the model is based on the composite statistics of the T_{ON} and T_{OFF} periods of the very large database used for our analysis. These long and composite statistical samples for the T_{ON} and T_{OFF} periods were produced by merging all T_{ON} and T_{OFF} periods accumulated individually for all the available call samples, and these composite samples are then used for analysis.

4 Logarithmic analysis of the composite samples

Previous work [11, 19] concludes that the human perception of time with respect to talk spurts and silences is logarithmic in nature. In line with their observations, we now produce loglinear plots of the probability densities of the composite measurement statistics for the T_{ON} and T_{OFF} periods.

4.1 Analysis of T_{OFF} periods

In Fig. 8, the composite measured result for the T_{OFF} periods is shown. We clearly observe that the T_{OFF} periods show a tri-modal structure on the logarithmic scale. The results suggest that at least 67% of silence lengths are less than 100 ms, a characteristic which many models have failed to capture. We also observe that a good majority of the silences are close to 10 ms and are related to respiratory pauses [17]. To this tri-modal structure we fit a tri-modal Gaussian Mixture model also shown in Fig. 8.

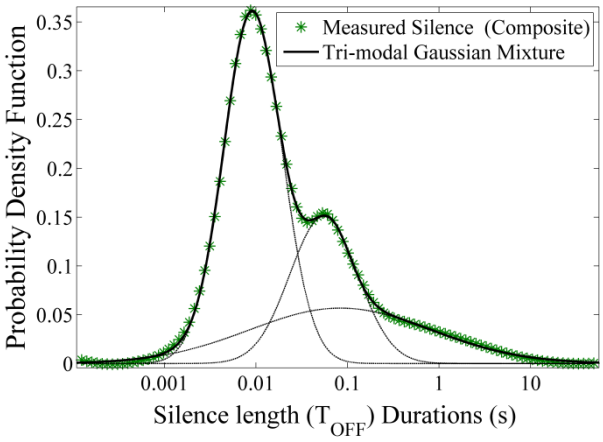

Fig. 8 *Tri-modal Gaussian profile for T_{OFF} periods*

The parameters of this Gaussian mixture model were estimated by way of the Expectation Maximization (EM) algorithm [20, 21], a procedure commonly used to estimate mathematical model parameters given that there is limited observed or incomplete data to complete the model. This Gaussian mixture model in the Logarithmic domain reflects a tri-modal lognormal mixture model in the linear domain for which the probability density equation is given by:

$$f_{T_{OFF}}(t) = \sum_{i=1}^{3} \alpha_i \frac{1}{t\ \sigma_i \sqrt{2\pi}} \exp\left(\frac{-(\log_{10}(t) - \mu_i)^2}{2\ \sigma_{i^2}}\right) \tag{2}$$

where α_i is the normalizing weight, μ_i is the mean and σ_i is the standard deviation of each Gaussian component representing the density of the T_{OFF} periods. The estimated values of these parameters are shown in Table 1.

4.2 Analysis of T_{ON} periods

When we repeat the same procedure as above for the T_{ON} periods, we observe a bi-modal profile for the density of T_{ON} periods in the logarithmic domain as shown in Fig. 9. Our results show that we have been able to accurately track talk spurts below 100 ms, which [7, 8] were not able to track and classify. We see that a significant number of the talk spurts which we have tracked lie in the range 6-10 ms in contrast to those longer than 100ms.

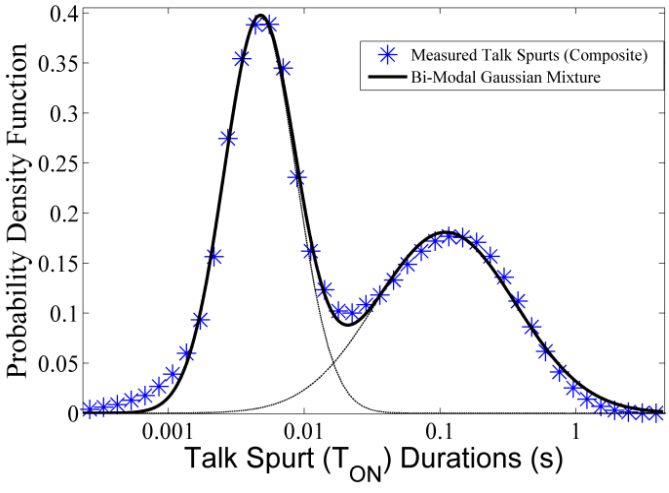

Fig. 9 *Bi-modal Gaussian profile for T_{ON} periods*

Whereas these spurts may represent elements of background noise, it is assumed, given we have T_{OFF} periods in the same range, that these small T_{ON} periods we detected are small vowel segments as pointed out in [17]. To this bi-modal structure we fit a bi-modal Gaussian mixture model as shown in Fig. 9 and again this reflects a bi-modal lognormal model in the linear domain, the density of which is represented by:

$$f_{T_{ON}}(t) = \sum_{i=1}^{2} \beta_i \frac{1}{t\ \varsigma_i \sqrt{2\pi}} \exp\left(\frac{-(\log_{10}(t) - \lambda_i)^2}{2\ \varsigma_{i^2}}\right) \tag{3}$$

where β_i is the normalizing weight, λ_i is the mean and ς_i is the standard deviation of each Gaussian component representing the density of T_{ON} periods and the estimated values of these parameters are also shown in Table 1.

Duration	***Parameters***	***Value***		
		i = 1	i = 2	i = 3
T_{OFF} (States 3,4,5)	α	0.56	0.11	0.33
	μ	-2.05	-1.29	-1.08
	σ	0.29	0.23	0.99
T_{ON} (States 1,2)	β	0.53	0.47	-
	λ	-4.653	-1.919	-
	ς	0.531	1.02	-

TABLE I
ESTIMATED VALUES OF PARAMETERS FOR LOGNORMAL DENSITIES FOR T_{ON} AND T_{OFF} PERIODS OF THE COMPOSITE LANGUAGE INDEPENDENT CALLS

5 On-off source model

Based on the fitted lognormal mixture distributions to the T_{ON} and T_{OFF} periods shown in the

previous section, a simple five state semi-Markov on-off model can be constructed as shown in Fig. 10. The three modes of the tri-modal lognormal distribution of the T_{OFF} periods in (2) are represented by the states 3, 4 and 5 respectively whereas the two modes of the bi-modal lognormal distribution of the T_{ON} periods in (3) are represented by the states 1 and 2 respectively. The states are connected by transitions between them as we now describe.

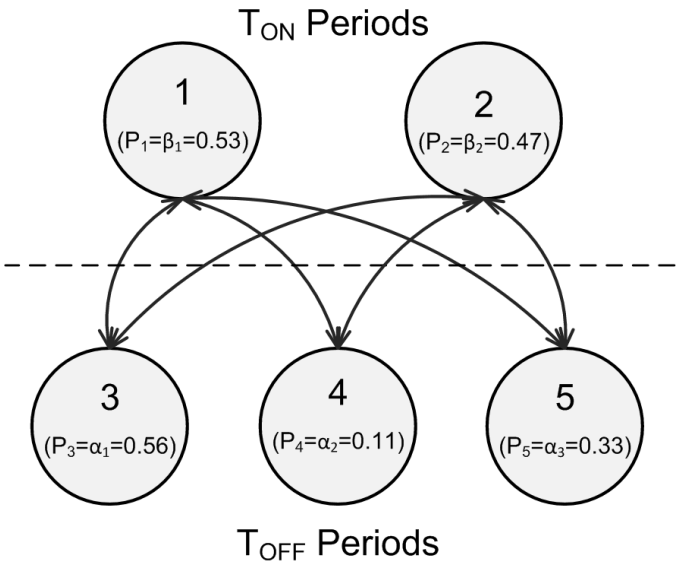

Fig. 10 *On-off semi-Markov source model*

The voice model transits between the ON and OFF states alternately as required and this reduces the number of potential transitions as compared to a general five state model. We also assume that the state transition probabilities are independent of the initial state which allows a unique assignment: the probability of visiting a state *i* is given by the normalizing weights α_i and β_i for the T_{ON} and T_{OFF} periods respectively, as shown in brackets in Fig. 10. This approximation is implicit in all previous work where correlations between ON and OFF periods are ignored. At any given moment the model exists in one of the five states. The time spent in each state is determined by the individual components of the complete tri-modal and bi-modal distributions (2) and (3) respectively. At the end of that time the model makes a transition, either from an OFF state to an ON state or vice versa. On leaving, for example, an OFF state the model makes a random choice according to the ON normalizing weights (β_i) as to which state to visit next so state 1 is chosen 53% of the time and state 2 is chosen 47% of the time. The parameters associated with the lognormal distribution depend on the mode each state represents and are shown in Table 1.

6 Simulated results and analysis

The proposed semi-Markov on-off model with lognormal transitions was implemented and Fig. 11 shows the CDFs for our measured and simulated T_{OFF}. We also compare these results to simulated results from [7] and measurements by [8]. As before, the CDFs for the T_{OFF} periods are represented in frames of 5 ms each. The figure clearly shows that our simulated results from the on-off model are in excellent agreement with our measured (composite) results for the T_{OFF} periods. In our results, we record an average of 30% more short silences than the English language based limited time measurements of [5] which used a frame by frame decision process for the silence detection. The figure also highlights the inaccuracy of the conventional 8-state model in tracking the T_{OFF} periods, especially for the smaller T_{OFF} periods.

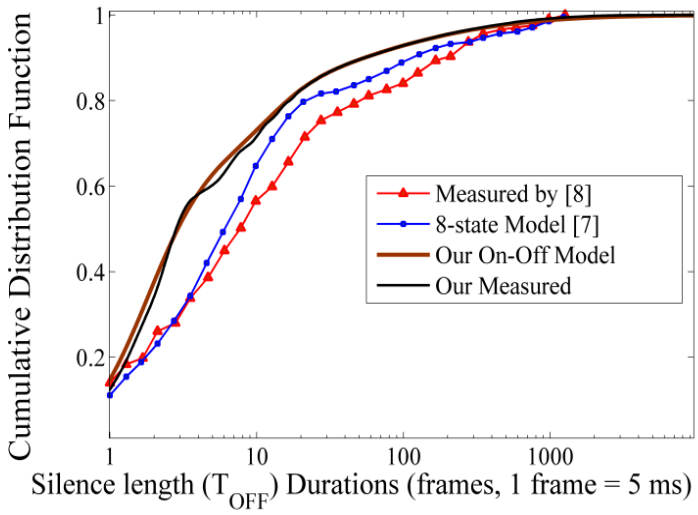

Fig. 11 *Cumulative distributions for T_{OFF} periods – measured values against semi–Markov model and results from [7, 8]*

Similarly, Fig. 12 shows the CDFs for the measured and simulated results of T_{ON} periods and those simulated by [7] and measured by [8].

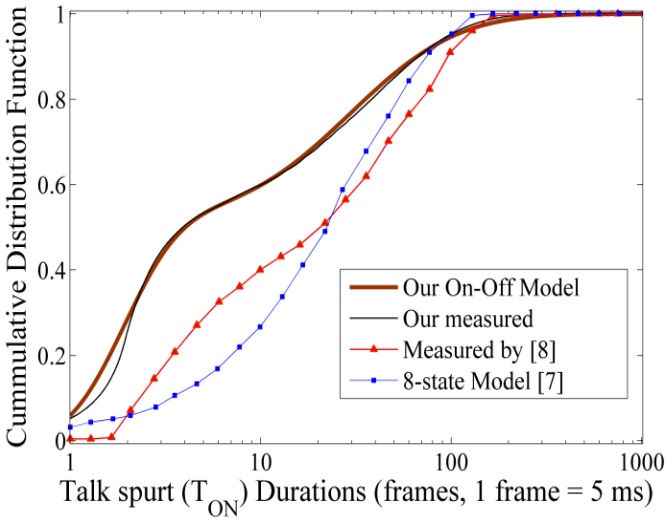

Fig. 12 *Cumulative distributions for T_{ON} periods – measured values against semi – Markov model and results from [4, 5]*

Again we observe that the simulated results of our lognormal model clearly fit our measured results. In contrast, we see that the results of the 8-state model widely differ with our results as a result of the lack of talk spurts shorter than 100 ms [7]. Also in our results we find that around 60% of the talk spurts for the multilingual database are below 50 ms in comparison to the 40% observed in [8]. Whereas it is also apparent from the figure, that around 25% of the talk spurts are below 10 ms, further analysis revealed that if we filter out all talk spurts below 10 ms from our composite sample, our measured result would look very similar to that of [8] as shown in Fig. 13. This highlights the inefficiency of the detecting methodology used by [8] to track short talk spurts, thus leading to the differences observed in our measurements.

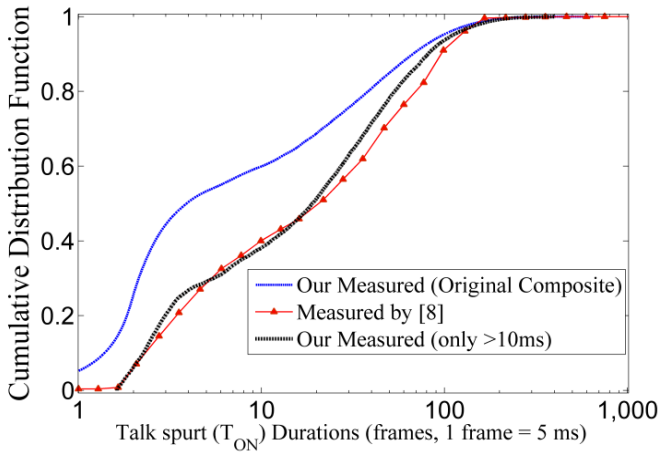

Fig. 13 *Cumulative Distributions for T_{OFF} periods – comparison of measured results of [8] against our measured results with all T_{OFF} < 10 ms excluded.*

7 Conclusions

We have presented accurate measurements of the on and off periods of VoIP activity from a large packetised multilingual database of two way call conversations and also from monologue speech data. This allowed a significant increase in the statistics compared to previous studies [7, 8]. Our study on the impact of linguistic differences on the VoIP call statistics revealed that the random dynamic nature of human interaction overshadows the linguistic differences, thus revealing the independence of the VoIP call statistics on language. The results presented here show that modelling VoIP calls solely based on the language and behaviour of an individual speaker would not lead to useful traffic models, thus modelling based on composite samples was performed and good average models of speaker and language independent statistics were derived. Further, the analysis of the composite samples revealed the heavy tailed and multi-modal lognormal nature of the T_{ON} and T_{OFF} periods.

We recorded a massive increase in short silence and talk spurts as compared to previous results. This suggests that these short silence periods can be well exploited by a sophisticated voice activity detector in terms of silence suppression with a new generation codec such as G.729b, to further enhance system capacity by increasing the efficiency of the traffic sources with increased OFF time. This is particularly crucial for wireless networks where spectral economy is vital. Conventional Markov source models with exponential on-off periods, still widely used because of ease of implementation, fail to accurately model our measurements. We conclude that our proposed lognormal on-off source model, which clearly tracks our measured results for T_{ON} and T_{OFF} very well and which can be easily implemented, should be used instead. In common with previous models, we assumed that the ON and OFF periods are uncorrelated. Any such correlations are likely to be language specific and since our model is an average over many languages, we believe that the approximation of independent statistics will be better in this case. Our measured and modelled results are of immediate use for further research on VoIP traffic modelling and on the choice and implementation of voice codecs supporting silence suppression.

8 Acknowledgements

This work was funded by the TSB under project HIPNET.

9 References

[1] Mockapetris, P.: 'Telephony's next act', *IEEE Spect.*, 2006, 43, pp. 28-32

[2] Menth, M., Binzenhofer, A., Muhleck, S.: 'Source models for speech traffic revisited', *IEEE-ACM Trans. Netw.*, 2009, **17**, (4), pp.1042-1051.

[3] Brady, P. T.: 'A model for generating on-off speech patterns in two-way conversations', *Bell Syst. Tech. J.*, 1969, **48**, pp.2445-2472

[4] Sriram, K., Whitt, W.: 'Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data', *IEEE J Sel. Areas. Commun.*, 1986, **4**, (6), pp. 833-846

[5] ITU-T: 'P.59: telephone quality objective measuring apparatus: artificial conversational speech', 1993

[6] Gruber, J. G.: 'A comparison of measured and calculated speech temporal parameters relevant to speech activity detection', *IEEE Trans. Commun.*, 1982, **30**, (4), pp. 728-738

[7] Stern, H. P., Mahmoud, S. A., Wong, K. K.: 'A model for generating on-off patterns in conversational speech, including short silence gaps and the effects of interaction between parties', *IEEE Trans. Veh. Technol.*, 1994, **43**, (4), pp.1094-1100

[8] Lee, H. H., Un, C. K.: 'A study of on-off characteristics of conversational speech', *IEEE Trans. Commun.*, 1986, **34**,. (6), pp. 630-637

[9] Dang, T. D., Sonokoly, B., Molnar, S.: 'Fractal analysis and modelling of VoIP traffic', Proc. Networks, Vienna, Austria, 2004, pp.123-130

[10] Chuah, C. N., Katz, R. H.: 'Characterizing packet audio streams from internet multimedia applications', Proc. IEEE ICC, New York, USA, 2002, pp. 1199-2203

[11] Casilari, E., Montes, H., Sandoval, F.: 'Modelling of voice traffic over IP networks', Proc. CSNDSP 2002, Staffordshire, U.K., 2002, pp.411-414

[12] Pragtong, P., Ahmed, K. M., Erke, T. J.: 'Analysis and modelling of voice over IP Traffic in the real network', *IEICE Trans. Inf. Syst.*, 2006, **E89-D**, (12), pp. 2886-2896

[13] Macwhinney, B.: 'The talkbank project', in Beal, J., Corrigan, K., Moisl, L (Eds.): 'Creating and digitizing language corpora: synchronic databases' (Palgrave-Macmillan, 2007), vol.1 [URL:http://talkbank.org/media/CABank/CallFriend/].

[14] Incorporated T., 'The open speech repository', [URL:http:/www.voiptroubleshooter.com/openspeech/index.html]

[15] Sanes, D. H., Reh, T. A. and Harris, W. A.: 'Behavioural development', in 'Development of the Nervous System' (Academic Press, 2006), pp. 303

[16] Goode, B.: 'Voice over internet protocol (VoIP)', *Proc. IEEE*, 2002, **90**, (9) , pp.1495-1517

[17] Zellner, B. 'Pauses and the temporal structure of speech', B. Bel, I. Marlin in 'Fundamentals of speech synthesis and speech recognition' (John Wiley, Chichester, 1994), pp. 41-62

[18] Rosner, B. S., Pickering, J. B.: 'Vowel constancy: coarticulation' E. Keller in 'Vowel Perception and Production' (Oxford University Press, 1994), pp.294-295

[19] Campione, E., Veronis, J.: 'A large-scale multilingual study of silent pause duration', Proc. Speech Prosody, Aix-en-Provence, France, Apr. 2002, pp.199-202

[20] Dempster, A.P., Laird N.M., Rubin D.B.: "Maximum likelihood from incomplete data via the EM Algorithm," *J. R. Stat. Soc.*, 1977, pp. 1-38

[21] Moon T. K., "The expectation maximization algorithm," *IEEE Signal Process. Mag.*, 1996, **13**, (6), pp. 47-60