Forced Alignment for Pronunciation Variation Analysis in Tatar

Matthias Drews

Yankı Öztürk

Abstract

One of the most significant challenges in large-scale analysis of recorded speech is the time consuming task of time alignment. This paper details the process of using Forced Alignment (FA) algorithms for automating such task, as well as subsequently evaluating the quality of and conducting phonetic analysis from resulting aligned data for Tatar, a low-resource language of the East European Plain.

1 Introduction

The task of large-scale phonetic analysis from recorded speech can be burdensome for linguists for many reasons, but among the most significant is the time-consuming alignment of transcription to audio. This type of processing, especially paired with the effort already involved in the transcription of large corpora, can significantly limit the amount of data researchers can consider for and include in their analysis. Automating this process of aligning transcription to audio is undoubtedly proves valuable in such cases.

Fortunately, advances in speech recognition technology resulted in language-specific acoustic models being sufficiently trainable with as little as an hour of transcribed data. (Johnson et al., 2018) This makes it possible for anyone to use these publicly available pre-trained models for the alignment of their field recordings.

In this paper, we detail the process of setting up a Forced Alignment (FA) pipeline for Tatar, an understudied language of the Turkic language family. We then discuss the quality of the results obtained with this alignment. In order to demonstrate the potential use of this technology in phonetic analysis, we perform speaker-level analysis of how the pronunciation of certain phones vary among the members of this community, based on societal factors like age and gender.

2 Background

2.1 The Tatar language

The Tatar language belongs to the Kipchak branch of the Turkic genus of the Altaic language family (Dryer and Haspelmath, 2013). It is spoken by around 5 million people in the autonomous Republic of Tatarstan region in west-central Russia, and many surrounding regions along the Volga river and in Western Siberia (Poppe, 1963).

There are three major dialects along with several mixed dialects of the Tatar language. The dialect of Kazan Tatars, also called the Central Dialect, forms the basis of the contemporary standard Tatar language, and is used in public offices, educational institutions, and media publications in Tatarstan (Burbiel, 2018). Tatar literary language uses the Cyrillic alphabet.

2.2 Phonetic analysis

The need for phonetic-acoustic analysis of recorded speech originates from the linguistic knowledge that the association between phonemes and the speech signal is not derived deterministically from a universal feature space, but is arrived at stochastically by learning generalizations across produced and perceived speech (Harrington, 2010).

Speech can be quantitatively measured as a sequence of pressure variations resulting from the particular movements of the articulatory tract during its production (Sarma and Prasanna, 2018). Phonetic analysis of speech signals is the task of identifying, distinguishing and segmenting these articulatory events that are reflected in the acoustic signal as phonetic segments.

2.3 Audio speech corpora

Recorded corpora of scripted, semi-spontaneous, or spontaneous speech, such as the Buckeye Corpus (Pitt et al., 2005) and TIMIT Continuous Speech Corpus (Garofolo et al., 1993) have long been com-

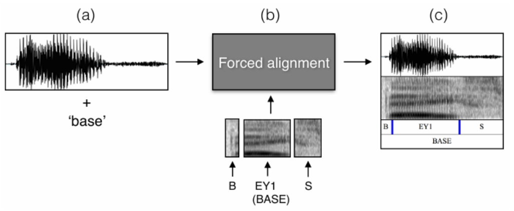


Figure 1: Diagrammatic representation of the Forced Alignment workflow. From *Automatic Detection of Sociolinguistic Variation Using Forced Alignment* (Bailey, 2016)

mon sources for conducting traditional linguistic research, and have been used more recently in developing speech recognition technologies.

Spoken language data can be prepared and read for recording, crowd-sourced from community members using their own recording devices, or come from traditional linguistic fieldwork sources (Chodroff et al., 2025). Various institutions such as the Bavarian Archive for Speech Signals at the University of Munich, and websites such as the Linguistic Data Consortium (Reed et al., 2008) make it their primary goal to make structured speech material available for speech and language technology.

A large corpus of speech data can be manually annotated in many different ways: among which may be orthographic transcription, phonetic transcription, part-of-speech tagging, lemmatization, and time-alignment. For large-scale phonetic analysis, full orthographic transcription of audio files is necessary (Olsen et al., 2017).

2.4 Forced alignment

In order to make use of corpus-based methods, there is a substantial need need for fast, inexpensive, accurate and consistent means of time-aligning arbitrary speech automatically (Wightman and Talkin, 1997). Forced Alignment (FA) is a family of algorithms for accomplishing this tedious task of placing time boundaries between words or phones in a speech recording. These algorithms take audio files, their orthographic transcriptions and a way to map graphemes to phonemes as input, and calculate time points that correspond to transitions (see Figure 1).

When compared to manual alignment, forced alignment is estimated to be at least twice as fast (Young and McGarrah, 2023). In addition to great savings on time and cost, with recent advancements, the reliability of forced alignment is considered almost equal to that of manual alignment by human experts (Adda-Decker and Snoeren, 2011). Consequently, forced alignment has become widely used not only in engineering text-to-speech systems, but also in research on sociolinguistics (Barth et al., 2020), historical linguistics and language change (Labov et al., 2013), psycholinguistics (Yuan et al., 2006), and language documentation and preservation (DiCanio et al., 2013).

2.4.1 Forced alignment systems

Aligners work by utilizing GMM-HMM architectures in their acoustic models. Each phone in a language's phonemic inventory, and each sequence of tri-phones in newer systems to allow for context sensitivity, is modeled as Hidden Markov Models with their acoustic properties as Gaussian Mixture Models.

Among the most widely used forced alignment tools by linguists are Montreal Forced Aligner (McAuliffe et al., 2017), MAUS (Schiel, 1999) and WebMAUS (Kisler et al., 2017), Prosodylab Aligner (Gorman et al., 2011), FAVE (Rosenfelder et al., 2022), and easyAlign (Goldman, 2011).

2.4.2 Forced alignment for LRLs

Traditionally, a major obstacle in employing forced alignment, the language-specific acoustic models for which were trained using up to hundreds of hours of speech data, has been the unavailability of such data for low-resource languages. It was previously assumed that a full alignment system could not be created for such languages (Kempton et al., 2011).

One method of overcoming this has been cross-language forced alignment, where aligners untrained on the target language are leveraged to perform alignment for under-resourced languages. Such systems have been reported as being suited to produce alignments of sufficient quality, with low disagreement rates with manually aligned data (Jones et al., 2019).

3 Experimental Setup

3.1 Data

We use version 23.0 of the Mozilla Common Voice dataset for Tatar (Ardila et al., 2020). The dataset includes a total of 30895 clips in various dialects of the Tatar language, making up 33 hours of recorded speech from 285 speakers.

The dataset download comes with `.mp3` audio clips and `.tsv` metadata. Each row of the metadata file represents a single audio clip, and contains a

gender	percentage
undefined	21.0
female	3.0
male	76.0

Table 1: Gender distribution in data

age band	percentage
undefined	21.0
twenties	5.0
thirties	71.0
fifties and older	3.0

Table 2: Age distribution in data

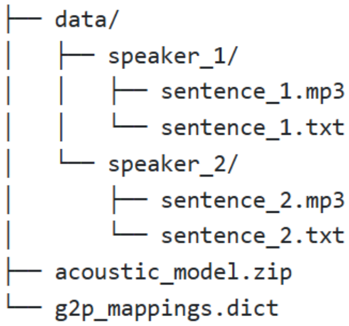
hashed ID of the speaker, relative path of the audio file, orthographic transcription of the audio, age of the speaker, gender of the speaker, and accent of the speaker in accordingly named columns.

Due to the crowd-sourced nature of Common Voice's creation, the demographic information in the dataset (see Table 1 and 2) are self-reported. The utterances are scripted and read aloud by contributors recording themselves using either the Common Voice website or phone application. Recordings are then validated by a different set of contributors.

3.2 Alignment pipeline

In preparation for the task of alignment, we sample 3000 instances from the test split of the dataset, extract corresponding orthographic transcriptions from its metadata table, and create our data directory (see Figure 2).

We use the Montreal Forced Aligner (McAuliffe et al., 2017), chosen for the availability of a downloadable pre-trained acoustic model and grapheme-to-phoneme dictionary for the Tatar language as part of the system. The alignment is carried out using the `mfa align` command included in the MFA suite. The results of alignment are exported automatically as Praat (Boersma and Heuven, 2001) `TextGrids` to their specified destination.



```
├── data/
│   ├── speaker_1/
│   │   ├── sentence_1.mp3
│   │   └── sentence_1.txt
│   └── speaker_2/
│       ├── sentence_2.mp3
│       └── sentence_2.txt
├── acoustic_model.zip
└── g2p_mappings.dict
```

Figure 2: Structure of an example directory for forced alignment with Montreal Forced Aligner

4 Results

4.1 Alignment analysis

In order to measure the accuracy of the alignment, we manually align 20 randomly sampled clips across all speakers, that are used as gold standard alignments. The evaluation measure is the proportion of generated time boundaries that were marked outside a 20 ms threshold of their counterparts in gold standard annotations, which is the most rigid threshold in forced alignment literature (Jones et al., 2019). We also report the mean and median error of the absolute timing errors on phone level (Table 3.

accuracy (%)	mean (ms)	median (ms)
38	773	16

Table 3: Alignment quality as measured by accuracy, mean error, and median error.

4.2 Linguistic analysis

The biggest task included in linguistic analysis is the qualitative comparison of linguistic phenomena by different demographic groups contained in the dataset. We analyze the average vowel, consonant and bigram length for each *age*, *accent* and *gender* group. Additionally, we include the `Textgrid` files for which no demographic information were stated in the corresponding `tsv` file.

The *ages* were categorized in ranges of 10 years, from the teens (10-19) until the seventies (70-79). *Accents* present are the Kazan and Mishar accents, both of which are present in various regions and republics of the Russian Federation. The stated *genders* of the speakers were `female_feminine` and `male_masculine`.

Each of these demographics are then analyzed by using the TextGridTools Python library (Buschmeier and Włodarczak, 2013). To do so, we group each IPA symbol into either consonants or vowels, and calculate the average length of vowels and consonants, as well as the average length of a mixed bigram, with the results presented in the following sections. Summarizing tables of length can be found Tables 5 and 4 in the Appendix.

4.2.1 Bigrams

Accent (Figure 3): The average Kazan bigram was 0.251 seconds, which is significantly longer than the average Mishar bigram at 0.189 seconds. The latter is much closer to the average bigram without accent marking, or ‘unmarked’ bigram at 0.182 seconds. There were very few accent marked results, with only 2 Kazan speakers and one Mishar speaker.

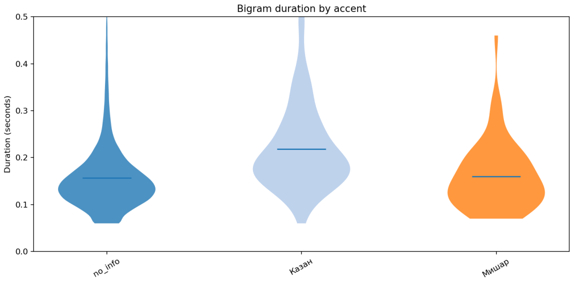


Figure 3: Bigram duration by accent

Age (Figure 4): The average duration of bigram length by age is fluctuating, starting at 0.217s for the teens, going down to 0.168s and 0.172s for the twenties and thirties respectively, reaching a peak at the fourties with 0.243, and going back down with 0.2s and 0.177 for the fifties and seventies. The unmarked bigrams had an average of 0.186s. The entries for fourties and seventies had only one entry.

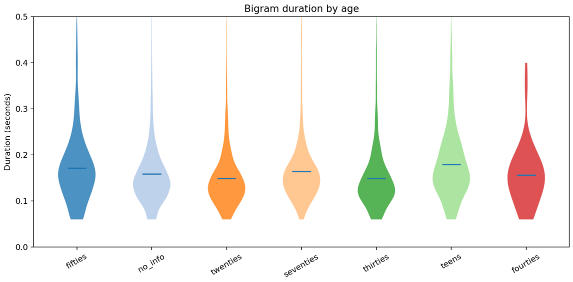


Figure 4: Bigram duration by age

Gender (Figure 5): The female and the unmarked participants were almost the same, with 0.185s and 0.186s respectively, with the male participants slightly faster at 0.172.

4.2.2 Consonants

Accent (Figure 6): Unmarked accents averaged out at 0.103 seconds per consonant, while the two Kazan speakers had 0.126 seconds, and the Mishar speaker 0.115s.

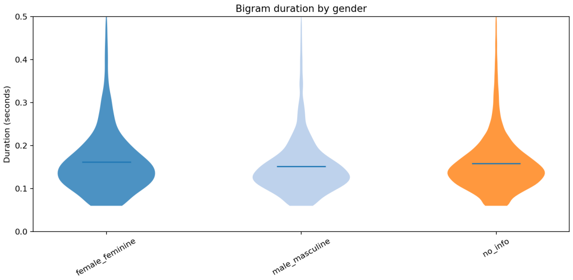


Figure 5: Bigram duration by gender

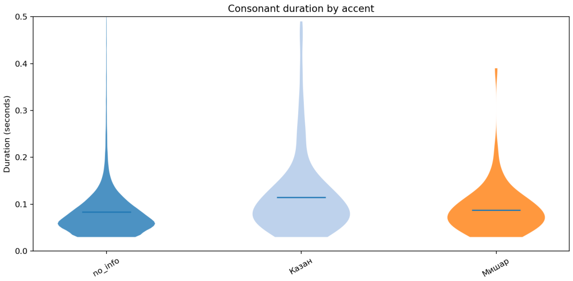


Figure 6: Consonant duration by accent

Age (Figure 7): Teens and fifties shared a length of 0.118, while twenties, thirties and seventies were close with 0.096, 0.094 and 0.094 respectively. The one speaker in their fourties had an average of 0.130, while the unmarked speakers took an average of 0.106 seconds.

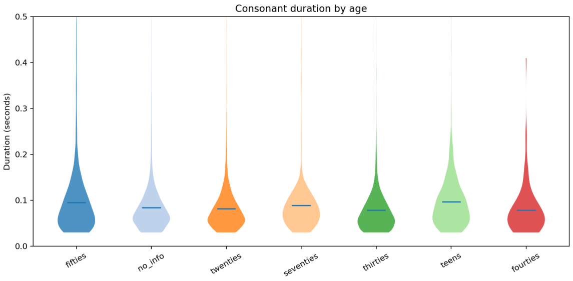


Figure 7: Consonant duration by age

Gender (Figure 8): Similar to bigrams, the female and the unmarked participants were almost the same, with 0.105s and 0.106 seconds, the males again slightly faster at 0.096s.

4.2.3 Vowels

Accent (Figure 9): The average vowel length for unmarked speakers was at 0.095 seconds, with the Kazan speakers at 0.139 and the Mishar speaker at 0.113s.

Age (Figure 10): The four speakers in their teens had the longest vowel length at 0.117s, with the fifties behind at 0.104s. The shortest duration was

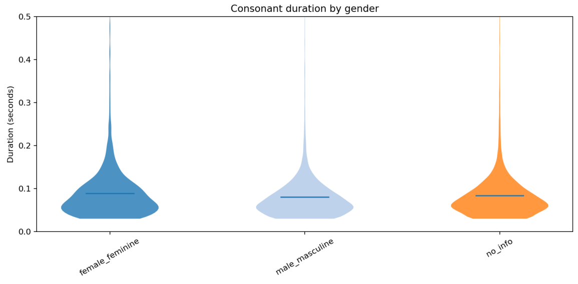


Figure 8: Consonant duration by gender

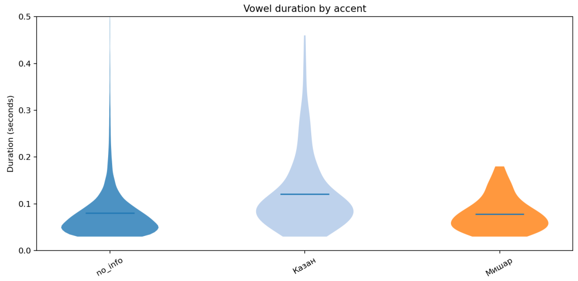


Figure 9: Vowel duration by accent

by the twenties, at 0.087s and the thirties at 0.091s. The fourties, seventies and unmarked were very close together at 0.097, 0.096 and 0.097s.

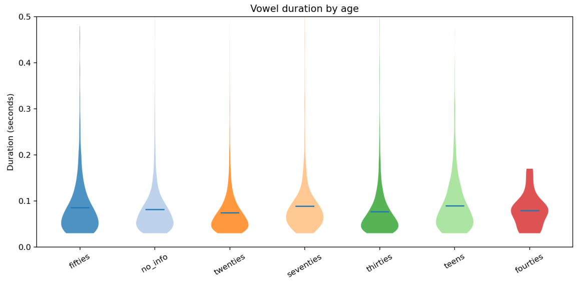


Figure 10: Vowel duration by age

Gender (Figure 11): Once again, the female speakers at 0.096 were extremely close to the unmarked speakers at 0.097 seconds, with the males at 0.091s only slightly faster.

4.2.4 Discussion

It should be noted that the number of speakers per demographic option vary wildly, ranging from only one person to several hundred. A larger and more balanced dataset is required to mitigate this problem.

While the analysis suffers from the small amount of speakers in some demographics, we can still make some interesting observations over patterns that hold true over all linguistic aspects. For example, in every analysis by gender, the female results are almost identical to the unmarked participants, while the males are slightly faster. The accent-marked participants had results that differed wildly from the unmarked participants, both faster and slower. Lastly, the age range had a fluctuating pattern, with the twenties and thirties always closer to each other compared to teens and fourties, as well as having usually the shortest durations.

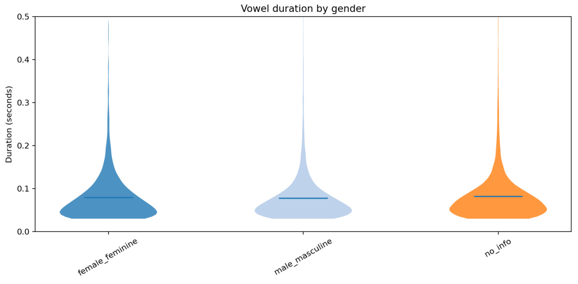


Figure 11: Vowel duration by gender

While there were some outliers, as visible in Figures 3 to 11, usually every consonant, vowel and bigram was between 0.1 and 0.5 seconds of length. Further exploration with more data would most probably yield clearer patterns, which would enable the predictive demographic profiling of an unmarked annotators.

5 Conclusion

In this paper, we detailed the process of setting up a Forced Alignment (FA) pipeline for Tatar, and reported our findings from performing phonetic analysis on 3000 force-aligned voice clips from a publicly available dataset. Our findings suggest that, while subtle, demographic groups of Tatar speakers tend to display predictable patterns that can, with more effort beyond our paper's scope, be used for predictive demographic profiling of individual speakers.

References

Martine Adda-Decker and Natalie D. Snoeren. 2011. Quantifying temporal speech reduction in french using forced speech alignment. *Journal of Phonetics*, 39(3):261–270.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

George Bailey. 2016. Automatic detection of sociolinguistic variation using forced alignment. *University of Pennsylvania Working Papers in Linguistics*, 22.

Danielle Barth, James Grama, Simon Gonzalez, and Catherine Travis. 2020. Using forced alignment for sociophonetic research on a minority language. *University of Pennsylvania Working Papers in Linguistics*, 25(2).

Paul Boersma and Vincent Van Heuven. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Gustav Burbiel. 2018. *Tatar Grammar*, 2 edition. Institute for Bible Translation, Stockholm.

Hendrik Buschmeier and Marcin Włodarczak. 2013. TextGridTools: A TextGrid processing and analysis toolkit for Python. In *Proceedings der 24. Konferenz zur elektronischen Sprachsignalverarbeitung*, pages 152–157, Bielefeld, Germany.

Eleanor Chodroff, Emily P. Ahn, and Hossep Dolatian. 2025. Comparing language-specific and cross-language acoustic models for low-resource phonetic forced alignment. *Language Documentation & Conservation*, 19:201–223.

Christian DiCanio, Hosung Nam, Douglas H. Whalen, H. Timothy Bunnell, Jonathan D. Amith, and Rey Castillo García. 2013. Using automatic alignment to analyze endangered language data: testing the viability of untrained alignment. *The Journal of the Acoustic Society of America*, 134(3):2235–2246.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. 1993. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium.

Jean-Philippe Goldman. 2011. Easyalign: an automatic phonetic alignment tool under praat. In *Interspeech 2011*, pages 3233–3236.

Kyle Gorman, Jonathan Howell, and Michael Wagner. 2011. Prosodylab-aligner: A tool for forced alignment of laboratory speech. In *Journal of the Canadian Acoustical Association*, volume 39.

Jonathan Harrington. 2010. *Phonetic Analysis of Speech Corpora*. Wiley Publishing.

Lisa M. Johnson, Marianna Di Paolo, and Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing prosodylab-aligner with tongan data. *Language Documentation & Conservation*, 12:80–123.

Caroline Jones, Weicong Li, Andre Almeida, and Amit German. 2019. Evaluating cross-linguistic forced alignment of conversational data in north australian kriol, an under-resourced language. *Language Documentation & Conservation*, 13:281–299.

Timothy Kempton, Roger Moore, and Thomas Hain. 2011. Cross-language phone recognition when the target language phoneme inventory is not known. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 3165–3168.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347.

William Labov, Ingrid Rosenfelder, and Josef Fruehwald. 2013. One hundred years of sound change in philadelphia: Linear incrementation, reversal, and reanalysis. *Language*, 89:30–65.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Rachel M. Olsen, Michael L. Olsen, Joseph A. Stanley, Margaret E. L. Renwick, and William Kretzschmar. 2017. Methods for transcription and forced alignment of a legacy speech corpus. *Proceedings of Meetings on Acoustics*, 30.

Mark A. Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

Nicholas Poppe. 1963. *Tatar Manual*, 1 edition. Indiana University Publications, Bloomington.

Marian Reed, Denise DiPersio, and Christopher Cieri. 2008. The Linguistic Data Consortium member survey: Purpose, execution and results. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ingrid Rosenfelder, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hillary Prichard, Jiahong Yuan, and Christian Brickhouse. 2022. Fave: Forced alignment and vowel extraction: v2.0.0.

Biswajit Dev Sarma and S. R. Mahadeva Prasanna. 2018. Acoustic–phonetic analysis for speech recognition: A review. *IETE Technical Review*, 35(3):305–327.

Florian Schiel. 1999. Automatic phonetic transcription of non-prompted speech. In *Proceedings of the XIVth International Congress of Phonetic Sciences*.

Colin W. Wightman and David T. Talkin. 1997. *The Aligner: Text-to-Speech Alignment Using Markov Models*, pages 313–323. Springer New York, New York, NY.

Nathan J. Young and Michael McGarrah. 2023. Forced alignment for nordic languages: Rapidly constructing a high-quality prototype. *Nordic Journal of Linguistics*, 46(1):105–131.

Jiahong Yuan, Mark Liberman, and Christopher Cieri. 2006. Towards an integrated understanding of speaking rate in conversation. In *Interspeech 2006*, pages paper 1795–Mon3A3O.1.

6 Appendix

Gender	Bigram	Consonant	Vowel
Female	0.185	0.105	0.096
Male	0.172	0.096	0.091
No Info	0.186	0.106	0.097

Table 4: Lengths by gender, in seconds

Age	Bigram	Consonant	Vowel
Teens	0.217	0.118	0.117
Twenties	0.168	0.096	0.087
Thirties	0.172	0.094	0.091
Fourties	0.243	0.130	0.097
Fifties	0.2	0.118	0.104
Seventies	0.177	0.094	0.096
No Info	0.186	0.106	0.097

Table 5: Lengths by age, in seconds