LEVERAGING ACOUSTIC INFORMATION FOR ENHANCED PERSONALIZATION IN THE ENTERTAINMENT DOMAIN

Charlie Bonfield, Manik Malhotra

TiVo, United States

ABSTRACT

The use of voice search has seen a significant increase over the past few years with the rise of voice-enabled devices. Voice search, by construction, affords information about the user that is not available in conventional text search. Most notably, implicit information obtained from raw audio can be used to tailor the underlying content retrieval system to more closely match user preferences. To maximize utility with minimal user input, however, an optimal voice search system should be able to perform tasks of this nature with minimal supervision. In this paper, we present a set of novel methods for inferring information about users of voice search without explicit enrolment and demonstrate subsequent enhancements to personalization. Further, we show how this work helps in reducing computational cost by reducing the number of possibilities considered by our natural-language understanding (NLU) system.

INTRODUCTION

Due to the rising popularity of virtual assistants like Amazon's Alexa, Apple's Siri, and Google Assistant, the modern consumer has grown accustomed to using conversation services while interacting with electronic devices and moving around the home to achieve tasks that would otherwise require "hands-on" human interaction, such as typing a query into a search engine or changing a music playlist. In addition to making these tasks simpler, the use of voice could provide additional contextual information (age, gender, sentiment, etc.) about the speaker that may then be further used to enhance the user experience.

In this paper, we describe a novel, efficient and effective strategy that could be used to enhance the discovery experience for voice remote users by attaching context to the string of text passed to backend search. The basis of the solution centres around personalization at the level of an anonymous but individual user in a household. To present our strategy, we have broken the end-to-end solution into multiple parts. First, we obtain information about the speaker using certain acoustic features extracted from the audio. Next, we fold in metadata that we have crafted around the offered content. This component is crucial for inferring the relevance of content to the individual user. In the scope of our research, this was limited to age-based relevance. The third step is to utilize what we have gleaned from the previous steps in real time to optimize natural language understanding and backend search. In doing so, we close the loop of enhanced entertainment discovery by matching a user to a collection of relevant content - in the absence of explicit user identification.

In what follows, we provide a discussion of the steps outlined above, including an overview of the dataset used during development, details related to the underlying learning algorithms used, and various measures of performance. We complement this work with results from case studies and explain operational benefits from our approach. In closing, we summarize our work and highlight areas of future research.

DATASET

Child Detection

To perform child detection, we used a random sample of audio files from an internal dataset representing a large population of anonymous households over a nine-month period. Using human annotators, we assigned each audio file in our sample a label indicating whether the speaker was an adult male, adult female, or child. This labelling, in turn, enabled us to apply supervised learning to our problem.

Child-Optimized NLU

We randomly sampled the above set of audio files to create a set of queries spoken by kids. Each audio file was transcribed using two leading ASRs and manually annotated to capture the exact spoken utterance. The output of each ASR was then passed through two NLUs: a baseline NLU and one with a custom “kids mode” enabled. The output was recorded as a combination of involved entities and the intent. This, in turn, lead to four sets of entity-intent combinations for each utterance. The manual annotations were analysed by a human and categorized in entity-intent by hand, resulting in one set of expected results for each utterance. The resulting observation and expectations are shown in Table 1.

Table 1: Observations and Expectations

ASR	NLU
ASR1	Baseline
ASR1	Special “kids mode” enabled.
ASR2	Baseline
ASR2	Special “kids mode” enabled.
Manual Annotation	Handpicked Entity/Intent for Kid

Figure 1: Frozen Movie Properties


```
"Frozen": {
    "Title": "Frozen",
    "Type": "movie",
    "ReleaseYear": 2013,
    "Ratings": ["U","FSK: 0", "Tous", "G", "6", "普遍級(普)"]
    "Genres": ["Animation", "Comedy", "Music", "Fantasy"],
    "Language": "eng",
    "Image":
"https://upload.wikimedia.org/wikipedia/en/thumb/0/05/Frozen_%282
013_film%29_poster.jpg/220px-
Frozen_%282013_film%29_poster.jpg"
    "Duration": 108}
```


DEVELOPMENT

Child Detection

Feature Extraction

To infer user demographic from raw audio, we used a combination of features motivated by previous works and handcrafted those which were shown to be significant for prediction. These features can be organized into a number of different classes (Table 2).

Table 2 - Taxonomy of features used for prediction

Feature Class	Examples	References
Mel-frequency cepstral coefficients (MFCCs)	MFCCs, deltas, double deltas (mean, stddev)	Tiwari (1)
Harmonics	Hand-crafted (total harmonic distortion)	-
Pitch	Fundamental frequency (f0), jitter	Boersma (2), Farrús et al (3)
Intensity	Intensity/loudness, shimmer	Farrús et al (3)
Speech Rate	Voiced-to-unvoiced ratio, estimated number of syllables/pauses	De Jong and Wempe (4)
Datetime	Time of day, day of week, weekday/weekend	-

Model Training

For prediction, we utilized an ensemble approach with a combination of linear and nonlinear classifiers (‘Zhou (5)’, ‘Chen and Guestrin (6)’) operating on the set of features discussed in the previous section. In our work, we found that such an approach was optimal given the nature of the expected class boundaries (Figure 1). More importantly, however, we were able to generate online predictions within 100 ms (on average) after receiving the audio file in its entirety, thereby allowing the response to be passed to the NLU without introducing latency that would adversely impact the user experience.

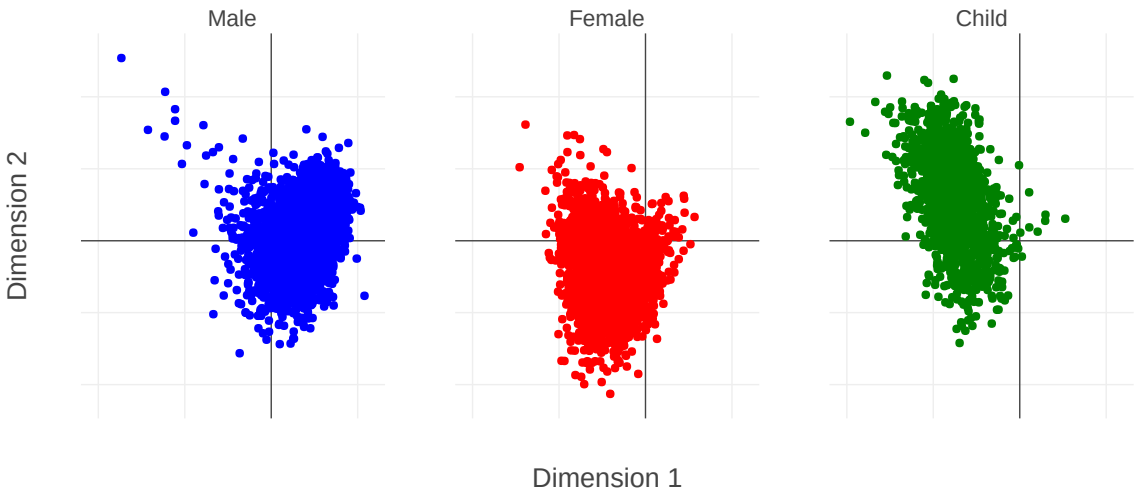

Figure 2 - Low-dimensional representation of acoustic features, highlighted by class.

Model Evaluation

To evaluate performance, we chose to examine the results using our folds from cross-validation due to the material and operational costs of annotating additional data to be

used as a holdout set. In particular, we paid attention to the following performance measures during development:

- Accuracy: (TP + TN) / (TP + TN + FP + FN)
- Precision (PPV): TP / (TP + FP)
- Sensitivity: TP / (TP + FN)
- Specificity: TN / (TN + FP)

where TP, TN, FP, and FN are the numbers of true positives/negatives and false positives/negatives, respectively.

Precision was of particular relevance for our study, as we wanted to be most confident in passing the child flag to the NLU. Since our method was inherently probabilistic, we also chose to introduce an optional pair of hyperparameters that served as thresholds that could be tuned in order to achieve an optimal balance between business use-cases and performance measures detailed above. Through the use of our ensemble model and additional hyperparameters, we were able to exceed 90% precision for the task of identifying children from audio files alone.

Metadata Tagging

Our strategy for metadata tagging is built around a voice search system. Specifically, it is designed for an ontology-based conversational question answering system. Systems such as these are built to answer questions from users in a conversational manner. They differ from legacy systems in that their knowledge of concepts is separate from language understanding. This is usually done by mining multiple sources - primarily encyclopaedias and catalogues – for relevant phrases, facts, and relations about content using Named Entity Recognition (Nothman et al (7)). For example, information about the movie *Frozen* may be mined from Wikipedia and stored as a JSON object, as shown in Figure 1.

The effort associated with the mining process can vary greatly in detail – for example, it could be as simple as tagging type and title or as complicated as tagging important plot keywords like “good vs evil” or “adoptive family”. For the scope of this paper, we limit the discussion to tagging age relevance.

Kids/Adults Labelling

For our work, we represent age relevance as a label. We labelled each video asset in the catalogue as either KIDS, ADULT, or UNKNOWN. This approach would result in *Frozen* being labelled as KIDS, whereas *Fifty Shades of Grey* would be labelled ADULT. We arrived at our set of three distinct labels by first categorizing each genre as either kid or adult friendly. For example, genres such as “Animation” or “Family” would be designated as kid friendly, whereas “Erotic” or “Slasher” would be adult friendly.

After genre labelling, we calculated the affinity of each asset towards kid and adult friendly genres. Genre importance was calculated using a simple term frequency-inverse document frequency (tf-idf) weighting, which caused popular genres like action, comedy, and drama to become irrelevant. The affinity was then further adjusted based on the ratings assigned to the movie. For example, a “U” (Universal) rating would further establish the affinity towards kid friendly and a rating of “R” (Restricted) would do the opposite. This leads to a label and a

score for that label. Due to the imbalance of genres with respect to age affinity, the three classes output from our system have different score ranges. All have a minimum value of 0, below which an asset is likely to belong to other classes. The UNKNOWN class is effectively unbounded because most of the genres can't be explicitly labelled as either kid or adult friendly, while the ADULT and KIDS classes each have a maximum score of 1000. In the ADULT class, higher scores are generally assigned to adult (explicit) films, whereas in the case of KIDS, shows for small children like *Peppa Pig* or *Teletubbies* score the highest.

For further understanding, consider a request where a user says "breadwinner". Since most NLUs perform an initial cleaning of text (via stemming and stop word removal), requests for the movie *The Breadwinner* and the show *Breadwinners* will be reduced to "breadwinner". This, in turn, leads to a single string of text being associated with two distinctly different assets - the movie, which is about war in Afghanistan, and the show, which is about two easy-going ducks. Our system tags the movie as UNKNOWN and the show as KIDS. This is mostly because the movie has an irrelevant genre and ratings that contradict the genres, whereas the show has multiple kid friendly genres and appropriate ratings. Refer to Figure 4 and Figure 5 for properties of each entity. With additional information relating to demographic, however, we would be able to more readily return the asset most relevant to the user. For additional examples of movies and labels, refer to Table 3.


```
"Breadwinner": {
    "Title": "The Breadwinner",
    "Type": "movie",
    "ReleaseYear": 2017,
    "Ratings": ["PG13","PG","Tous","G"]
    "Genres": ["Drama", "Animation", "War"],
    "Language": "eng",
    "Image": "https://upload.wikimedia.org/wikipedia/en/7/78/The_Breadwinner_%28film%29_poster.jpg"
    "Duration": 93
}
```


Figure 3: *The Breadwinner*, UNKNOWN (1100)


```
"Breadwinner": {
    "Title": "Breadwinners",
    "Type": "tvseries",
    "ReleaseYear": 2014,
    "Ratings": ["6+","TVY7", "G", "6", "T"]
    "Genres": ["Animation", "Comedy", "Animated Comedy", "Children"],
    "Language": "eng",
    "Image": "https://upload.wikimedia.org/wikipedia/en/thumb/1/1a/Breadwinners_%28TV_series%29.svg/800px-Breadwinners_%28TV_series%29.svg.png"
    "Duration": 22}
```


Figure 4: *Breadwinners*, KIDS (309)

Table 3: Examples of movies with their associated labels/scores.

Movie	Class (Score)
Shutter Island	ADULT (631)
BlacKkKlansman	ADULT (619)
White Boy Rick	ADULT (577)
A League of Their Own	UNKNOWN (1535)
Fantastic Beasts and Where to Find Them	UNKNOWN (1397)

The Blind Side	UNKNOWN (1389)
Moana	KIDS (392)
Shrek 2	KIDS (352)
The Good Dinosaur	KIDS (232)

APPLICATIONS

Discussion of study related to Kids NLU

In order to develop a kid-optimized NLU, we chose to not follow the traditional approach of supervised learning. Primarily, we did so due to the limited amount of data we had available for this task – since we did not have a large amount of data available for training, we instead used a small amount of data during development. These cases were analysed in great depth to understand voice search usage amongst kids. When we went to test the NLU, however, we considered about 1000 cases. The main goal of this analysis was to understand the link between typical queries we would expect to observe in the field and the expected result. In practice, this can be particularly difficult since the interaction between kids and technology is much more natural compared to adults. This, in turn, required subjective analysis to understand the logic between observation and expectation. For example, if a kid were to say, “Minions Minions and Planet of the Apes”, deciding on an acceptable response would require a great deal of subjectivity.

Our study of these cases leads to a lot of interesting observations, the most important of which are as follows:

Pronunciation
Kids tend to mumble a lot and still expect the system to understand them. This is very different from an adult who would usually respond to a system failure by speaking clearly. Kids, on the other hand, rarely correct their pronunciation or rate of speaking.

Poor Reconstruction
Kids try to reconstruct longer, harder-to-remember phrases by replacing some words with either imitations of their pronunciation or random noises. For example, instead of saying “Transformers: Robots in Disguise” a kid may just say “Formers: Bots Disguise” while filling the missing parts with sound imitating their pronunciation. This obviously leads to erroneous transcription from the ASR.

No Clear Intention
Many times, kids do not have a clear intention when making a voice search. They are much more likely to play with the system than adults. With that in mind, there is a lot of guessing involved to understand what exactly their expectation is from a voice command.

In light of the observations detailed above, we decided to make the following changes:

Adopting a Stricter Approach to NER
To adopt a more binary approach to named-entity recognition (NER), we were far more flexible when it came to recognizing entities with high kid’s affinity and very strict towards

entities with high adult affinity. With our system, we organized this flexibility into two categories: phonetic variance and normal variance. Phonetic variance solves the pronunciation problem - this corrects errors where the ASR comes up with a similarly sounding but completely different (in terms of direct string comparison) transcriptions. Normal variance, on the other hand, is for handling missing words. In our NLU, we solve this with a fuzzy chunker that can be programmed to be more tolerant of missing words/phrases when looking up known phrases.

Adding Implicit Bias towards Certain Entities

We optimized our existing entity selection algorithm which was earlier defined in Sashikumar et al (8), by introducing a bias in our entity scoring function which was passed to the conditional random field (CRF). Previously, this function was primarily matching the phrase to entity type weights, for example, according to our ontology-based knowledge system that was optimized for the entertainment domain. “Mahatma Gandhi”, for example, is 46.8% likely for entity type “Person” and 48.8% for “Movie”. In essence, the function was ranking different types and passing the rank-ordered list as a feature to the CRF model. For example, in the mentioned case, we would end up sending two features - *type1*: Movie, *type2*: Person.

Based on our findings, we modified the scoring function to introduce an additional conditional penalty motivated by the fact that our existing system was biased towards adults, as most of the existing training data was defined by adults. Modifying the existing scoring function allowed us to use our existing model without retraining. Instead of simply boosting “kid friendly” entities, we introduced a cost sensitive function which penalized entities conditionally based on the demographic information from raw audio. This function considered three factors: demographic information from the raw audio, entity type, and normalized age relevance. The optimal function introduced high penalties in cases where we detected a kid was speaking, and the age relevance was high for adults. However, the function introduced very little penalty in opposite cases (i.e. an adult searching for an entity with high relevance for kids).

$$F(demography, phrase, entity) = F_{ex}(phrase, entity) - \sum_{d=0}^{k} a_d \times rel_d(D_d, entity) \times (1 - sim(demography, D_d))$$

D_d = Demography class, for example in our case $D_0$was ADULT, D_d KID, and D_2, UNKNOWN.

F_{ex}= Existing scoring function for scoring relevance from phrase to entity.

a_d= Configurable coefficient for each D_d.

rel_d= Relevance function optimized for each demography.

sim = A similarity function for two demography’s, max is 1 when D_d is the current user demography and min is 0.

Results

To compare the baseline NLU with the optimized NLU, we tested it on both our existing test set (19,000 queries, primarily from adults) as well as the 1,000 cases defined earlier. For the existing test set, we assumed that the demography was always unknown and only compared the NLU, thus ignoring the ASR completely. This allowed us to examine any adverse effects that the optimized NLU might have introduced for existing users, which we found to be insignificant (< 0.05% affected). For the other 1,000 cases, we compared the output with Manual Annotation, with the results presented in Table 4:

Table 4: NLU Testing Results

ASR	NLU	Passed Cases	Net Accuracy	Gain
ASR1	Baseline	451/1000	45.1%	N/A
ASR1	Special “kids mode” enabled.	513/1000	51.3%	+6.2%
ASR2	Baseline	871/1000	87.1%	+42%
ASR2	Special “kids mode” enabled.	992/1000	92.2%	+5.1%

Integration with Backend Search

The scope of this study was limited to benefits of demography detection for speech recognition and natural language understanding. The underlying motivation for this work centred around the fact that kids generally directly ask for content instead of trying to discover something, as adults do. That aside, the user experience can also be further improved by folding these features into a content recommendation system. This would apply to cases where either the user showed an intention to discover (a broad request for “movies”, for example) or where the requested content isn’t available and one wants to recommend available assets relevant to the user instead.

On a broader scale, demographic detection holds particular promise for both collaborative- and content-based recommendation. For the former, this information would be most useful to have for the purposes of filtering or boosting assets at query time. For content-based recommendation, this feature fits naturally within the respective vector spaces. Instead of depending on recommender to tease apart clusters of KIDS, ADULT, and UNKNOWN assets (which would likely be emergent from genres), one can directly fold in this information. The same reasoning applies to user demography, where rather than asking the user to explicitly enter information or inferring demography from the nature of the requested content alone, one can infer it from acoustic features.

CONCLUSION

In summary, we have presented an end-to-end method for enhanced personalization via voice search for the entertainment domain. Through acoustic feature extraction, we are able to infer demographic information about users, which in turn enables us to make use of augmented metadata and an optimized NLU during the process of content retrieval.

Looking to the future, there is a wealth of opportunities available for this work to be extended. Approaches that fall within the realm of zero-shot learning, as discussed in Wang et al (9), when utilized in the context of speaker identification provides a natural

mechanism for linking an utterance to an individual user within a household without enrolment. This, in turn, allows for even greater advances in personalization that may be achieved seamlessly without explicit action required from the user. On the backend side, one could imagine leveraging information captured at the level of the individual to further limit the breadth of interpretations from the NLU and allow for a far more optimized search. These avenues, amongst others, remain in active development and hold the potential to provide a truly personalized content discovery experience.

REFERENCES

1. Tiwari, V. 2010. MFCC and its applications in speaker recognition. International Journal on Emerging Technologies 2010, 1 (1). pp. 19 to 22.

2. Boersma, P. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proceedings of the Institute of Phonetic Sciences 1993, 17. pp. 97–110.

3. Farrús, M., Hernando, J., Ejarque, P. 2007. Jitter and Shimmer Measurements for Speaker Recognition. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2007, 2. pp. 1153 to 1156.

4. De Jong, N.H. and Wempe, T. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. Behavior Research Methods. 2009, 41 (2). pp. 385 to 390.

5. Zhou, Z. H. 2012. Ensemble Methods: Foundations and Algorithms.

6. Chen, T. and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785-794.

7. Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R. 2013. Learning multilingual named entity recognition from Wikipedia. Artificial Intelligence. 2013, 194. pp. 151–175.

8. Venkataraman, S. and Mohaideen, N. 2015. A Natural Language Interface for Search and Recommendations of Digital Entertainment Media.

9. Wang, W., Zheng, V. W., Yu, H., and Miao, C. 2019. A Survey of Zero-Shot Learning: Settings, Methods, and Applications. ACM Transactions on Intelligent Systems and Technology (TIST) February 2019. 10 (2).