INFORMATION-THEORETIC CRITERIA FOR UNIT SELECTION SYNTHESIS[1]

Jon Yi and James Glass

Spoken Language Systems Group
MIT Laboratory for Computer Science
Cambridge, Massachusetts 02139 USA
{jonyi,glass}@mit.edu

ABSTRACT

In our recent work on concatenative speech synthesis, we have devised an efficient, graph-based search to perform unit selection given symbolic information. By encapsulating concatenation and substitution costs defined at the class level, the graph expands only linearly with respect to corpus size. To date, these costs were manually tuned over pre-specified classes, which was a knowledge-intensive engineering process. In this research paper, we turn to information-theoretic metrics for automatically learning the costs from data. These costs can be analyzed in a minimum description length (MDL) framework. The performance of these automatically determined weights is compared against that of manually tuned weights in a perceptual evaluation.

1. INTRODUCTION

The recent surge in popularity of concatenative methods [2, 5, 6, 10] for TTS can be attributed to many factors including more storage, faster computation, and, perhaps most importantly, increased naturalness. We believe that, in the context of spoken dialogue systems, the speech synthesizer should place as little listening burden on the user as possible. As conversational systems deployed over the telephone are displayless by nature, it is especially important to have natural and intelligible speech generation to reduce the user's cognitive load. In previous work, we have demonstrated that this is an achievable task in limited domains using word and sub-word units [17, 18].

Because synthesis researchers believe that signal processing should be minimized to maintain quality [5], success in the earlier stage of unit selection, where speech segments are selected from a database according to an input specification, becomes vitally crucial. We have taken the optimization formulation that decouples concatenation and substitution costs (which describe where and what to join) and cast it into a finite-state transducer (FST) framework [13]. By only describing costs at the class level (not at the instance level), we obtain transducers whose topology scales linearly with respect to corpus size. In application domains where the vocabulary and grammar is known ahead of time, a synthesis corpus can be designed around the requirements of the task. What remains to be addressed is the formulation of the synthesis costs and classes.

The approach we take in this work is to assume that the members of a defined contextual equivalence class are deemed to have similar acoustic effects. Furthermore, certain equivalence classes should be more similar than others to allow for relaxation of constraints, or backoff. If all alternatives should be equally poor, then the classes are not defined sufficiently distinct. What we have described here is the familiar notion of the substitution cost. When enough data is observed for each class, we can build statistical models that summarize the nature of each class. These models can be compared with models of other classes to determine the substitution costs. Next, we treat the notion of the concatenation cost. What is desired is a metric that will describe how easily two speech segments can be joined without significant perceptual distortion. Observations can be collected around concatenation boundaries and joint statistics summarizing behavior around those boundaries can be calculated. We shall see later how information-theoretic measures can be used to define suitable substitution and concatenation costs.

Much of the earlier work in the literature has concentrated on instance-level costs that directly compare speech segments, or instantiations of the speech units. Numerical metrics such as Euclidean, Kullback-Leibler, and Mahalanobis distances calculated over spectral features have been considered [3, 4, 10–12, 16]. Because we define costs at the class level for scalability, we apply the metrics not to pairs of individual examples but to pairs of distributions of multiple examples.

2. REVIEW OF UNIT SELECTION

Unit selection involves finding an appropriate sequence of units, $\hat{u}$, from a speech corpus given an input specification, u. Because the process can be formulated as a search, what is appropriate is determined by minimizing a pre-determined search metric:

$$\hat{u} = \underset{\hat{u}}{\operatorname{argmin}}\, J(u, \hat{u})$$

While the units and specification can encompass information from many linguistic levels, in the rest of this paper we shall primarily focus on phone-sized units. The sequence of units that has minimal cost is used in a subsequent process of waveform generation.

Hunt and Black [10] cast the problem of unit selection as a constrained optimization problem with two types of costs: a concatenation cost which describes the quality of the join between two

[1] This research was supported by DARPA under Contract N66001-99-1-8904 monitored through Naval Command, Control and Ocean Surveillance Center, and contract DAAN02-98-K-003, monitored through U.S. Army Natick Research Development and Engineering Center. The first author is a recipient of an Intel Corporation Fellowship.

In *Proceedings of the 7th International Conference on Spoken Language Processing*, Sep. 16-20, 2002, Denver, Colorado, pp. 2617-2620.

units, and a substitution cost which describes degradation in using a unit from a context different from the specification. When two segments are contiguous in the speech database, their concatenation cost is defined to be zero. When contexts match exactly, the substitution cost is also zero. This permits the greedy selection of variable-length units or non-uniform units [14]. In a later section we shall propose a method for automatically determining the numerical value of the costs from data.

2.1. MDL FORMULATION

Now we consider a communication-theoretic formulation of the unit selection process in which the unit selector is a black box receiving an input specification and transmitting the best match and associated waveform segments. This black box can be thought of as a noisy transmission channel, because the output may not precisely match the input. Corruption of the message is quantified by description lengths which describe the penalty in approximating the input with the output. When a message is transmitted through the unit selector without degradation, then it must be that the desired units are entirely contiguous within the speech database.

In Figure 1, the input specification, u, passes through unit selection and exits as $\hat{u}$. For the purpose of this discussion, we ignore the associated waveform segment descriptors on the output side and do not display the side information provided by the synthesis costs and corpus. The input and output are streams of interleaved unit and transition symbols. The unit symbols may contain contextual (e.g., triphone $\beta(\alpha : \gamma)$ denotes β in the context of α and γ) information as well. Transition symbols are pseudo-units that have no acoustic realization but serve as place-holders between units. Transition symbols are marked differently whether they bridge two contiguous units ($\alpha|\beta$) or not ($\alpha\#\beta$). Naturally, all transition symbols on the input side are marked as contiguous ($\alpha|\beta$) because that is the ideal case.

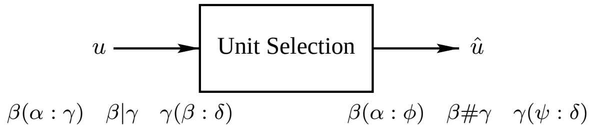

Fig. 1. Communication-theoretic formulation of unit selection.

In unit selection, the process of determining the best match is managed by a set of substitution (S) and concatenation costs (C). We use the two costs to quantify the degradation in unit and transition symbols, respectively. To handle the pathological cases described above, S is zero when the input and output symbols are identical and C is zero when a contiguous transition symbol appears unchanged at the output side. For the remaining cases, in the next section we turn to costs automatically determined from data. For now, turning back to Figure 1, we see that the total cost, $J(u, \hat{u})$, for the example inputs and outputs is:

$$J(u, \hat{u}) = S_{\beta(\alpha:\gamma)\rightarrow\beta(\alpha:\phi)} + C_{\beta\#\gamma} + S_{\gamma(\beta:\delta)\rightarrow\gamma(\psi:\delta)}$$

Since we are minimizing a sum of description lengths, we refer to this setup as a minimum description length (MDL) formulation. To combat quadratic growth in parameters, an independence assumption is made and substitution costs for triphones are decoupled into left-sided and right-sided components. That is, when $\beta(\hat{\alpha} : \hat{\gamma})$ is used in place of $\beta(\alpha : \gamma)$, the total substitution cost is:

$$S_{\beta(\alpha:\gamma)\rightarrow\beta(\hat{\alpha}:\hat{\gamma})} = S^{l}_{[\alpha]\beta\rightarrow[\hat{\alpha}]\beta} + S^{r}_{\beta[\gamma]\rightarrow\beta[\hat{\gamma}]}$$

3. ACOUSTIC MODELLING

In this section we describe a framework for modelling acoustical observations within and across units. Although we base our modelling on the acoustic front-end of the SUMMIT speech recognizer [8], these ideas should generalize to any model when consistently applied. One distinguishing aspect of the currently proposed method is that the observation space is the same for models within and across units. This permits a more direct comparison of concatenation and substitution costs.

As depicted in Figure 2, measurements, x, are made at boundaries between α and β, which may be speech units, or, more generally, classes of speech units. The notation, $\alpha[\beta]$, refers to a unit, α, with β on its right side. Similarly, the notation, $[\alpha]\beta$, refers to a unit, β, with α on its left side. As in the SUMMIT system, we form the observation vector from a telescoping average of mel-frequency cepstral coefficients on either side of the boundary [7].

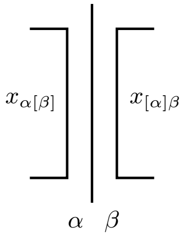

Fig. 2. Graphical representation of boundary measurements.

As we will see in the next section, we collect second-order statistics (i.e., mean and covariance) on measurements observed across boundaries. Note that with the above definitions, joint statistics ($x_{\alpha|\beta}$) can be collected in one pass and split into prospective and retrospective statistics. The mean vector is halved and the upper left and lower right blocks of the covariance matrix are pared off. The prospective observation space ($x_{\alpha[\beta]}$) will be used to define the right-sided substitution costs and the retrospective observation space ($x_{[\alpha]\beta}$) will be used to define the left-sided costs.

4. INFORMATION-THEORETIC DISTANCE METRICS

In creating an automatic procedure for determining synthesis costs, we have used the Kullback-Leibler (KL) divergence measure as our main workhorse. Although KL distance, $\mathcal{D}(p \,||\, q)$, has many interpretations in different contexts, the underlying theme is that it represents the asymmetric cost of approximating p with q.

Because Kullback-Leibler distance captures the notion of asymmetric substitution and provides an associated cost, we use it for defining substitution costs. As we shall see, the KL distance between contexts which have similar acoustical effects will be low. Note that the right-sided substitution cost looks forward in time and that the left-sided substitution cost looks backward in time.

$$\begin{aligned} S^{r}_{\alpha[\beta]\rightarrow\alpha[\gamma]} &\equiv \mathcal{D}\left(p(x \mid \alpha[\beta]) \,||\, p(x \mid \alpha[\gamma])\right) \\ S^{l}_{[\beta]\alpha\rightarrow[\gamma]\alpha} &\equiv \mathcal{D}\left(p(x \mid [\beta]\alpha) \,||\, p(x \mid [\gamma]\alpha)\right) \end{aligned}$$

Another information-theoretic measure closely related to KL distance is mutual information which measures statistical independence between two random variables. It can be defined as the

KL distance between a joint distribution and the product of its marginal distributions. Another definition relates mutual information to the difference between the unconditional and the conditional entropies. If the mutual information is low across a concatenation boundary, the conditional entropy is high and little information is communicated across the boundary. An information impasse makes for a good concatenation point.

$$C_{\alpha\#\beta} \equiv \mathcal{I}\left(p(x \mid \alpha[\beta])\ ;\ p(x \mid [\alpha]\beta)\right)$$

For multivariate Gaussian random variables, expressions in terms of nats (natural bits or $\ln 2$ bits) for entropy, mutual information, and Kullback-Leibler distance are:

$$\begin{aligned} \mathcal{I}(P\ ;\ Q) &= \tfrac{1}{2}\ln\left(\tfrac{|\Sigma_{pp}|\ |\Sigma_{qq}|}{|\Sigma|}\right),\ \Sigma = \left[\begin{array}{cc} \Sigma_{pp} & \Sigma_{pq} \\ \Sigma_{qp} & \Sigma_{qq} \end{array}\right] \\ \mathcal{D}(P\,||\,Q) &= \tfrac{1}{2}\left((\mu_Q - \mu_P)^T \Sigma_Q^{-1}(\mu_Q - \mu_P)\ + \right. \\ &\qquad\qquad \left. tr(\Sigma_P \Sigma_Q^{-1} - I) - \ln|\Sigma_P \Sigma_Q^{-1}|\right) \end{aligned}$$

We model the observations with Gaussian distributions because closed-form expressions exist for their entropies. Although others have used numerical integration or empirical estimates [15], Gaussian distributions are simple to estimate with only first- and second-order statistics needed. Furthermore, the distribution for a class of units can be formed by combining statistics of individual units in a bottom-up fashion. Finally, when the observation space for substitution and concatenation costs is the same, the distributions can be more directly compared. As a result, the entropies and, thus, the synthesis costs are of the same scale. However, it is still permissible to use a weighting factor to trade off between substitutions and concatenations. This weight will mostly likely depend on listener preference. Since the synthesis costs are based on non-negative quantities, the ramification for applicability to Viterbi and A^* search algorithms is that using Viterbi partial path costs for upper-bound estimates in a subsequent A^*-based, N-best extraction step is admissible.

Because the information-theoretic criteria introduced here are generic, the effectiveness of the synthesis costs will mostly depend on the design of the observation space and subsequent density estimation. As in most speech recognition work, we have assumed normality or Gaussianity for simplicity. Although recognition and synthesis are inverse problems with many commonalities, the best performance in each problem may not necessarily be achieved by the same type of features. The current observation vector looks the same amount ahead and backward in time. On one hand it might be conceptually attractive to design heterogenous features [9]. For example, the time windows could adapt to the phonetic context, or they might be biased towards the future for modelling speech sounds in languages that have more anticipatory rather than carry-over co-articulation. On the other hand, there would be the issue of normalizing or balancing the costs as the observation spaces would now be different. We note that the concatenation cost defined here may bear similarity to splicing cost as defined in other work [4].

5. ANALYSIS

We are currently using this information-theoretic framework to automatically determine synthesis costs from the very synthesis corpus used in unit selection. As the synthesis costs are speaker-dependent, this represents a matched condition. The current synthesis corpus consists of around 100 minutes of in-house recordings of typical system responses in weather information, flight status, and air travel domains.

In the left part of Figure 3 we see a bubble plot of concatenation costs where the rows and columns represent the context to the left and the right of the concatenation boundary, respectively. The contexts are vowels, semivowels, nasals, and obstruents. For the purpose of visualization, we use the radius of a circle (not the area) to depict the magnitude of a cost. In the past we have hypothesized that places of source changes are more appropriate for concatenations. Indeed we see that concatenations between vowels and semivowels, for example, where the source does not change, are most costly. Semivowel-nasal and vowel-vowel concatenations are the next most costly. Fricative-vowel, nasal-stop, and fricative-semivowel boundaries are the least costly concatenations.

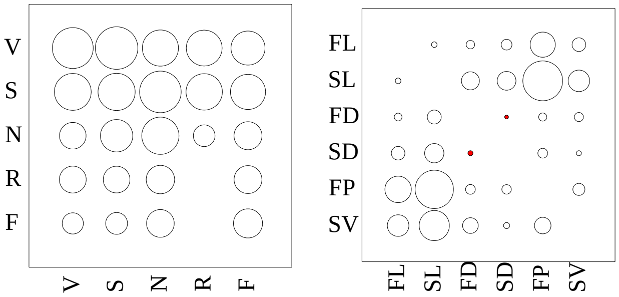

Fig. 3. Left: Bubble plot of concatenation costs matrix. Rows and columns correspond to context (vowel, semivowel, nasal, stop release, fricative) on the left and right sides of phonetic boundaries. Right: Bubble plot of substitution costs matrix for phonetic context to the right of vowels. Rows and columns correspond to true and approximating contexts (fricatives and stops in labial, dental, palatal, and velar places of articulation).

To understand substitution costs, we turn to the right part of Figure 3 which displays the right-sided costs for vowels in terms of varying contexts: fricatives and stops in labial, dental, palatal (more precisely, fricatives and affricates), and velar places of articulation. Because the lips are responsive articulators, transitions leading into labial stops are very spectrally distinctive and this is reflected in the large substitution costs seen in the second row. Large costs also appear for interchanging palatal and labial places. Finally, it is undesirable to substitute a velar stop with a labial stop.

Because the dynamic range of the concatenation costs does not appear to be high, it is possible that the concatenation classes are overly broad. In contrast, the classes in the example substitution cost matrix are low in size with only two or four members each. We have chosen to keep the classes that were manually designed in earlier work to better isolate experimental conditions. Although not reported here in depth, we are currently investigating the use of automatic decision trees for clustering concatenation and substitution contexts. For example, phone-specific classes for substitution costs are derived by successively splitting left and right contexts independently for all phones using phonological questions.

6. RESULTS

In the 2001 DARPA Communicator Evaluation, naive users were recruited from across the country and asked to call a flight travel system deployed by one of eight different participating sites. Approximately 20 subjects called each system four times each, and another 15-16 subjects called each system eight or more times.

After each call, users were asked to fill out a Likert scale questionnaire in which they were asked to rate, among other things, the degree to which they agreed with the statement, "I found the system easy to understand in this conversation." Each of the eight participating sites used some form of concatenative synthesis in their deployed systems. Among the eight, the MIT system, which used the ENVOICE synthesizer, was ranked the highest (4.07/5) in user agreement with that statement. Although this particular configuration concatenated words and phrases and used unit selection search more sparingly, it nonetheless provides a reference point for naturalness for the baseline system used in these experiments.

In order to evaluate the information-theoretic measures, we have conducted a small-scale listening evaluation on a test set of twenty-two utterances in an air travel domain. Using a leave-one-out approach, we re-synthesize an utterance from the remainder of the corpus using oracle phonological information. Concatenation boundaries are smoothed with windowed overlap-and-add processing. The shift corresponding to maximum correlation between abutting short-time spectral frames is used to appropriately offset the windows at the concatenation boundary. Since window shifts at boundaries introduce variable timing of the window centers, normalization is required to preserve energy relations.

In comparing synthesizers using manually tuned and automatically derived synthesis costs, we asked eleven subjects to perform A/B comparisons on an utterance-by-utterance basis. While only five out of eleven subjects found the second system preferable overall, pooling votes on individual utterances showed that utterances produced by the second system were found to be preferable 55% of the time, or roughly half the time. A 50% preference level suggests that subjects are indifferent and that one system is indistinguishable from the other. Since the baseline system is considered quite natural, we view this as a positive result as it reduces the amount of manual heuristics necessary to tune a synthesizer.

In the three utterances where the second system received one vote or less, spurious pitch and irregular durations were detracting factors. On the other end of the spectrum, three utterances received nine votes or more. Using automatically learned costs reduces the concatenation rate (i.e., number of concatenations per second of synthesized speech) from 2.09 to 1.77, a 15.3% relative reduction.

7. CONCLUSIONS & FUTURE WORK

In this paper, we have introduced a data-driven formulation based on information-theoretic measures for automatically determining the concatenation and substitution costs used in unit selection methods. Preliminary listening tests show that this automatic approach gives a favorable result while significantly reducing synthesizer development time. We are currently automating the design of equivalence classes in ongoing work. Future work includes evaluating these synthesis equivalence classes, developing an intonation model, and exploring hierarchical extensions to unit selection.

8. ACKNOWLEDGMENTS

Joe Polifroni and Stephanie Seneff helped in system development and in the DARPA evaluation. Michelle Spina recorded the corpus used in these experiments. Chao Wang provided the automatic decision tree tool along with suggestions of its application. Members of the Spoken Language Systems Group participated in listening evaluations and gave helpful feedback.

9. REFERENCES

[1] L. Baptist and S. Seneff, "GENESIS-II: A versatile system for language generation in conversational system applications," *Proc. ICSLP*, III:271–274, Beijing, 2000.

[2] M. Beutnagel, M. Mohri, and M. Riley, "Rapid unit selection from a large speech corpus for concatenative speech synthesis," *Proc. Eurospeech*, 607–610, Budapest, 1999.

[3] I. Bulyko and M. Ostendorf, "Joint prosody prediction and unit selection for concatenative speech synthesis," *Proc. ICASSP*, II:781–784, Salt Lake City, 2001.

[4] I. Bulyko and M. Ostendorf, "Unit selection for speech synthesis using splicing costs with weighted finite state transducers," *Proc. Eurospeech*, II:987–990, Aaalborg, Denmark, 2001.

[5] N. Campbell, "CHATR: A high-definition speech re-sequencing system," *Acoustical Society of America and Acoustical Society of Japan, Third Joint Meeting*, December 1996.

[6] A. Conkie, M. Beutnagel, A. Syrdal, and P. Brown, "Preselection of candidate units in a unit selection-based text-to-speech synthesis system," *Proc. ICSLP*, III:314–317, Beijing, 2000.

[7] J. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-based speech recognition," *Proc ICSLP*, IV:2277-2280, Philadelphia, 1996.

[8] J. Glass, T. J. Hazen, and L. Hetherington, "Real-time telephone-based speech recognition in the JUPITER domain," *Proc. ICASSP*, 61–64, Phoenix, 1999.

[9] A. Halberstadt, *Heterogenous Acoustic Measurements and Multiple Classifiers for Speech Recognition*. Ph.D. thesis, MIT, 1998.

[10] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proc. ICASSP*, Atlanta, 373–376, 1996.

[11] E. Klabbers and R. Veldhuis, "On the reduction of concatenation artefacts in diphone synthesis," *Proc. ICSLP*, 1983–1986, Sydney, 1998.

[12] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Transactions on Speech and Audio Processing*, 39–51, January 2001.

[13] E. Roche and Y. Shabes (eds.), "Finite-State Language Processing," MIT Press, 1997

[14] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," *Proc. ICASSP*, 679–682, New York, 1988.

[15] P. Viola, N. Schraudolph, and T. Sejnowski, "Empirical entropy manipulations for real-world problems," *Neural Information Processing Systems*, volume 8, MIT Press, 1996.

[16] J. Wouters and M. Macon, "Control of spectral dynamics in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, 30–38, January 2001.

[17] J. Yi and J. Glass, "Natural-sounding speech synthesis using variable-length units," *Proc. ICSLP*, 1167–1170, Sydney, 1998.

[18] J. Yi, J. Glass, and L. Hetherington, "A flexible, scalable finite-state transducer architecture for corpus-based concatenative speech synthesis," *Proc. ICSLP*, III:322–325, Beijing, 2000.